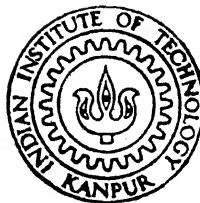


SEPARATION OF GLOTTAL WAVE AND VOCAL TRACT TRANSFER FUNCTIONS BY SUCCESSIVE ITERATION

By

Maj. TARANJIT SINGH



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
FEBRUARY 1991

EE
1991
M
SIN
SEP

SEPARATION OF GLOTTAL WAVE AND VOCAL TRACT TRANSFER FUNCTIONS BY SUCCESSIVE ITERATION

*A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of
MASTER OF TECHNOLOGY*

By

Maj. TARANJIT SINGH

to the

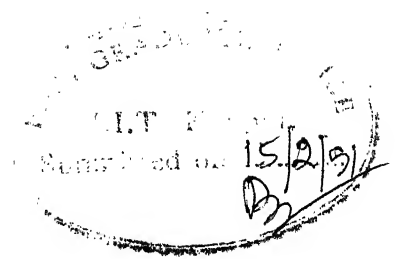
DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

FEBRUARY 1991

0 6 12 11 11

LIBRARY
112189

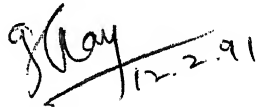
EE-1991-M-SIN-SEP



CERTIFICATE

This is to certify that the thesis entitled 'Separation of Glottal Wave and Vocal Tract Transfer Function by Successive Iteration' is a record of work carried out under my supervision by Maj Taranjit Singh and to best of my knowledge it has not been submitted elsewhere for a degree.

February, 1991


(Dr. G.C Ray)

Assistant Professor

Department Of Electrical Engineering

Indian Institute Of Technology

KANPUR 208016, INDIA

dedicated to
my mother Sardarni Gurdev Kaur, and
father Sr Sadhu Singh in their everlasting and loving memory
and
my wife Ravi
and
children Areet Kaur and J. Karan Singh

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to

– Dr. G C Ray, my thesis supervisor, for his excellent guidance, wholesome stimulus and encouragement and above all, his vigourous and considerate approach.

– Miss Seshu K, project associate who provided all the necessary help in my early part of hardware fabrication.

– my wife Ravi for her sacrifices and facing the hardships caused by my long absence from household chores help.

Maj Taranjit Singh.

SYNOPSIS

A speech signal $S_r(z)$ of a sustained vowel can be represented in frequency domain as the product of $P(z).G(z).V(z).R(z)$, where $P(z)$ is the transfer function of train of impulses, $G(z)$ is the glottal wave transfer function, $V(z)$ is the vocal tract transfer function and $R(z)$ is radiation load transfer function. Mathematically,

$$S_r(z) = P(z).G(z).V(z).R(z).$$

The radiation load component of speech signal is due to conversion of the volume velocity of sound coming out of lips into the pressure waves received at the microphone placed in distant field. This conversion is in the form of a differential relationship between volume velocity and sound pressure. To remove the effect of this component, a process called the Inverse Filtering using a digital integrator had been suggested by many authors but in the present case a conventional analog integrator was used whose performance is very similar to the digital integrator (of the form $1/(1-\alpha z^{-1})$) suggested by them. The speech signal then can be represented as

$$S(z) = P(z).G(z).V(z) \text{ or equivalently } S(z) = P(z).H(z)$$

$$\text{where } H(z) = G(z).V(z) \text{ and } S(z) = S_r(z)/R(z)$$

$P(z)$ component of $S(z)$ was separated out using a technique called Homomorphic Deconvolution. This removal of $P(z)$ was carried out in the laboratory using a FFT Analyser. In other words, $H(z)$ was recovered from $S(z)$ and was transferred to PC

through a GPIB interface.

The $G(z)$ and $V(z)$ components are found to be superimposed in frequency domain and are therefore difficult to separate. Their separation is what the thesis has endeavoured to achieve. L. R. Rabiner and Ronald W. Schafer showed that a lossless vocal tract system divided into N identical sections, can be characterised by a set of its area functions or, equivalently, reflection coefficients. Mathematically $V_a(z)$, the vocal tract transfer function is

$$V_a(z) = \frac{0.5 (1+r_0) \prod_{k=1}^N (1+r_k) z^{-N/2}}{1 - \sum_{k=1}^N \alpha_k z^{-k}}$$

Using this $V_a(z)$ the glottal wave transfer function $G_a(z)$ was obtained by simple mathematical division of $H(z)$ by $V_a(z)$ in the frequency domain. The $G_a(z)$ so obtained was represented by synthetic glottal wave transfer function and was utilized to separate the individual $V_i(z)$ from $H(z)$. By their successive iteration and using the relation between linear predictor coefficients and PARCORS, the area functions of the individual vocal tract were determined from $V_i(z)$. This helped in reconstruction of individual vocal tract. This reconstruction is of immense help in diagnosis of pathological disorders.

CONTENTS

	Page
Chapter-1: Introduction	1-27
Chapter-2: Homomorphic Deconvolution	28-41
Chapter-3: Successive Iteration	42-48
Chapter-4: Linear Predictive Coding	49-54
Chapter-5: Losses in Vocal Tract	55-61
Chapter-6: Conclusion and scope for further work	62-63
References	64-66

LIST OF SYMBOLS

$V_a[z]$	Average area vocal tract transfer function
$V_i[z]$	Individual area vocal tract transfer function
$G_a[z]$	Average area glottal wave transfer function
$G_i[z]$	Individual area glottal wave transfer function
$V[z]$	Vocal tract transfer function
$P[z]$	Excitation pulse transfer function
$R[z]$	Radiation load transfer function
$G[z]$	Glottal wave transfer function
$S_r[z]$	Transfer function of the complete speech model
$S[z]$	Transfer function speech model without radiation load
$H[z]$	Transfer function of speech model after removal of $P[z]$ and $R[z]$
$g[n]$	Impulse response of glottal filter

CHAPTER 1

INTRODUCTION

Speech is a peculiar human activity not endowed to other species. Speech is also a primary means of communication between people. In order to understand this activity it is essential to know the fundamentals of speech production process. Development of good digital processing technique has made the processing of speech signal in real time feasible. Specifically we shall be concerned with digital signal processing technique to study the steady state behaviour of the vocal tract system during the production of single sustained vowel [4].

However before the objective of the thesis is outlined, it is imperative that the process of speech production and mechanism of speech production is explained briefly .

1.1 THE PROCESS OF SPEECH PRODUCTION

Speech signals are composed of a sequence of sounds. These sounds and the transition between them serve as a symbolic representation of information. Most languages including English can be described by the distinctive bits of sounds called *phonemes*. In American english there are about 42 phonemes including vowels; diphthongs; semivowels; and consonants. Due to limit on the rate of physical motion

human articulators produce speech at the rate of 10 phonemes/sec. If we represent each phonemes by a set of six bit binary numbers average information rate will thus be 60 bits/sec [1,4].

AIM

Speech production mechanism suggests that if individual vocal tract parameters are extracted from speech signals, they could provide important advantages.

These are:

- (i) They provide useful clues about the defective vocal tract of a pathological source so that it can be diagnosed and treated for correct speech utterance.
- (ii) Identification of speaker for his secrecy.
- (iii) Automatic speaker recognition.

1.1.1 MECHANISM OF SPEECH PRODUCTION

Figs 1.1 and 1.2 show the important features of human mechanism that constitute production of speech signals [1,4]. The vocal tract begins at the opening between vocal ^hcords and the glottis and ends at lips. The vocal tract itself consists of pharynx (the connection from esophagus to the mouth) and the mouth. In an average adult male the total length of vocal ^hcord is 17 cm approximately. The cross sectional area is determined by the positions of tongue; lips; jaws; and velum. If divided into N equal sections the area of the vo^cal tract will vary

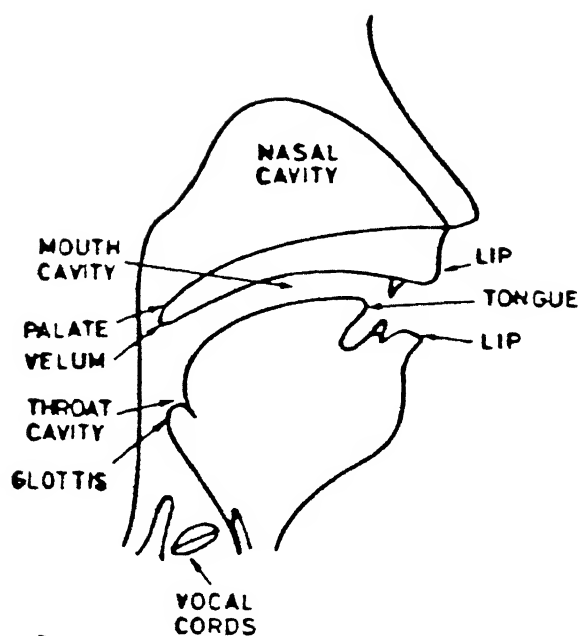


FIG. 1.1 SPEECH PRODUCTION MECHANISM

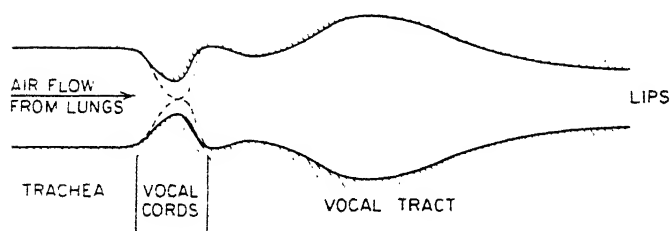


FIG 1.2 SCHEMATIC REPRESENTATION OF VOCAL SYSTEM.

at each section. We observe that vocal tract is a circular with differing areas. The area can vary from zero indicating complete closure to about 20 cm^2 . The nasal tract or nasal cavity begins at the velum and ends at the nostrils. When velum is lowered, the nasal tract is acoustically coupled to vocal tract to produce nasal sounds of speech.

The subglottal system consisting of lungs, bronchi and trachea, serves as the source of energy for production of speech. The speech is an acoustic wave that is radiated from this system when air is expelled from the lungs and resulting flow of air is perturbed by a constriction somewhere in the vocal tract. [1,2]

For getting the required quality of sound the vocal chords' tension must change. When these vocal chords which are muscle fibres, are tightly closed, pressure is built up below it. When certain threshold level of pressure, depending upon tensions in vocal chords and sufficient to force vocal chords to open, is built up, a narrow glottis opening is created between them. A jet of air rushes out with great kinetic energy and this due to Bernoulli's law causes the fall of potential energy (and hence pressure) in the glottis. Vocal chords will close again constricting the passage of air flow. This process is repeated 50-400 times in one second. Thus the vocal chords enter a state of sustained oscillations. The rate of opening

and closing of glottis is controlled by the air pressure in the lungs, the tension and stiffness of the vocal cords and the glottal opening under rest conditions.

1.1.2 Classification of speech

Speech can be classified into three distinct classes according to the mode of excitation of vocal tract. Specifically,

- *Voiced sounds* are produced by exciting the vocal tract with quasi periodic pulses of air flow caused by the opening and closing of glottis. The examples are: /u/, /d/, /i/, /e/.
- *Fricative/Unvoiced sounds* are produced by forming a constriction some where in the vocal tract and forcing the air through constriction so that turbulence is created, thereby producing a noise like excitation. Examples are /f/, /θ/, /sh/ etc.
- *Plosive sounds* are produced by completely closing off vocal tract, building up pressure behind the closure, and then abruptly releasing the pressure. Examples are /ts/ etc.

In each case speech signal is produced by exciting the vocal tract system with wide band excitation. The different sounds then produced depend on the the constriction of vocal tract which is a tube having different area functions in different sections along its

length. These areas vary from person to person. Also, vocal tract changes shape relatively slowly with time and this can be modelled as slowly time varying filter that imposes its frequency response properties on the spectrum of the excitation.

The utterance of a vowel is voiced sound. The variation of the area at each section along the length of the vocal tract determines the resonant frequencies or the formants of the speech signal. The dependence of cross sectional area upon the distance along the tract is called the area function of the vocal tract and it is this area function which determines the shape of the vocal tract [5].

1.2 MAKING A MODEL OF SPEECH

Having discussed briefly the mechanism of speech production, it will be better if we consider mathematical representations of speech production and make a model of speech signal.

1.2.1 Glottal Excitation

The glottal excitation for voiced sounds (in our case a sustained vowel) is the appropriate source for vocal tract excitation [16]. As brought out earlier, the rate of opening and closing of glottis is controlled by the pressure in lungs, the tension in vocal cords and the area of glottal opening under rest condition. In other words, we can say that the glottal excitation is a sequence or train of impulses which are spaced by the desired

fundamental period for a short interval segment of speech signal. A digitized model of glottal wave is shown in Fig 1.3. This wave shape has two parts namely, the open glottis having rising phase and the falling phase and closed glottis. The frequency response of such a wave shape for ideal values of n_2 and n_3 is also shown in Fig 1.4 .

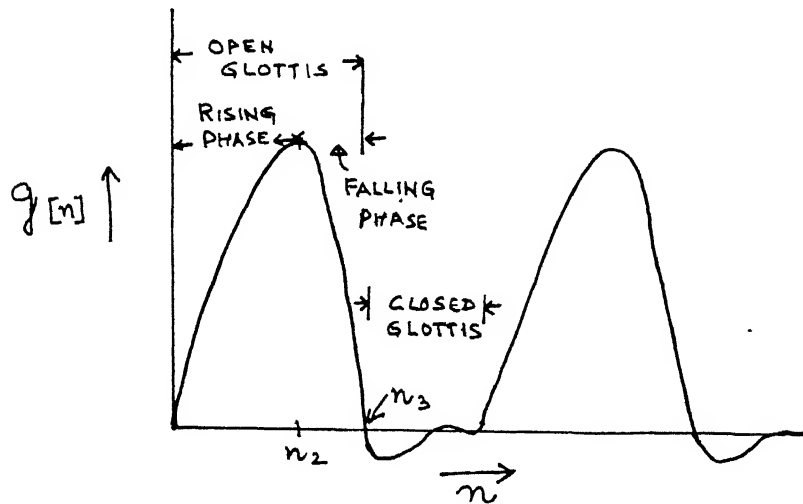


FIG 1.3 IDEAL GLOTTAL WAVE FORM.

Mathematically, each phase can be described as

- (a) Rising Phase of open glottis:

$$S_1[n] = \hat{S} * 0.5[1 - \cos \omega_R n], \quad n=0,1,2,\dots,n_2 \quad \dots 1.1$$

where ω_R the pulse rise frequency = π/n_2 .

- (b) Falling phase of open glottis:

$$S_2[n] = \hat{S} * [K \cdot \cos(\omega_R n - \pi) - K + 1] \quad \dots 1.2$$

where $n = n_2(1), n_2(2), \dots, n_3$ and

K =steepness factor, ideally $0.5 < K < 10,000$

K can be calculated from

$$n_3 - n_2 = \frac{1}{\omega_R} \cos^{-1} \left(\frac{K-1}{K} \right) \quad \dots 1.3$$

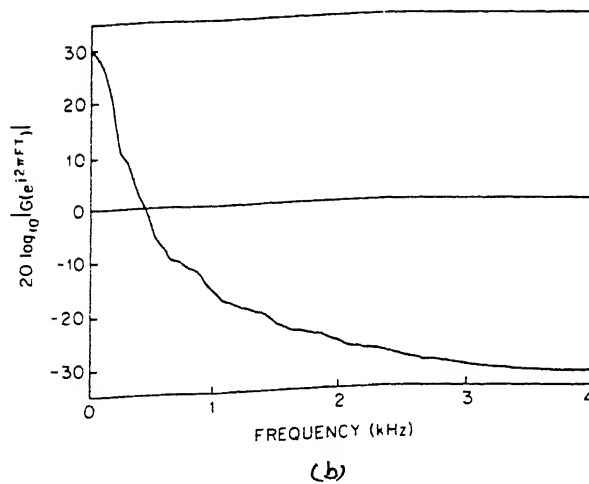
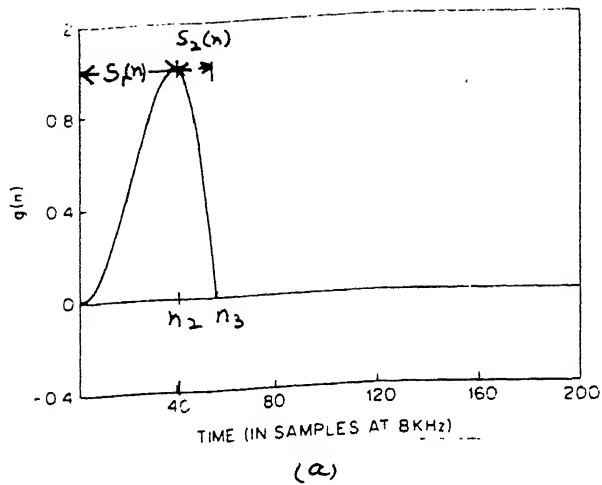


FIG 1.4(a) IDEAL GLOTTAL WAVE AND
(b) FREQUENCY RESPONSE OF (a)

(c) Closed glottis:

$$S_3[n] = 0.$$

n is the number of samples which characterize a modelled glottal wave shape.

The values of n_2 and n_3 had been particularly useful, as we shall see later (Chapter 3), in getting and shaping the individual glottal wave.

1.2.2 Excitation of vocal tract

As brought out earlier, the excitation of vocal tract is caused by glottal excitation. It has also been pointed out earlier that the vocal tract can be assumed to be a slowly varying filter that imposes its frequency response properties on the spectrum of glottal excitation. Mathematically therefore, $G(z)$ and $V(z)$ are superimposed on each other [7,21].

1.2.3 Terminal analog model

Having discussed the process of speech production mechanism in sufficient details, a simple source filter model as shown in Fig 1.5 can now be made. The vocal tract is represented by a time varying filter. Excitation source is a quasi periodic impulse train or excitation generator. Amplitude control regulates the energy output. Various parameters for vocal tract filter, voiced/unvoiced switch, pitch period and amplitude are regularly updated so as to

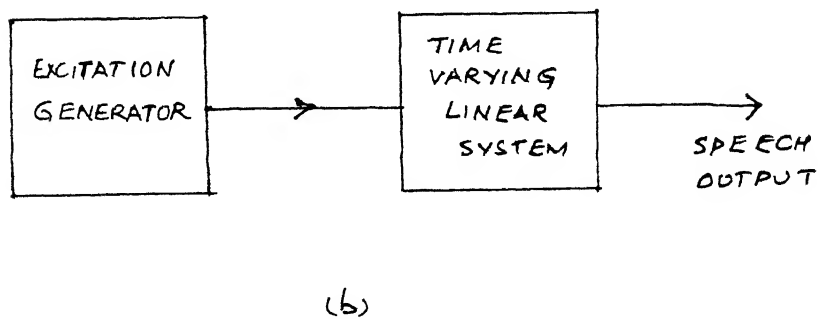
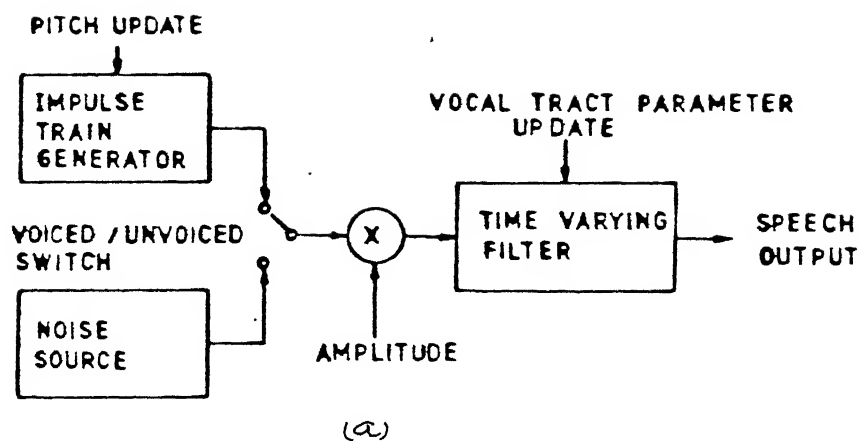


FIG 1.5(a) SOURCE FILTER MODEL, AND
(b) EQUIVALENT TERMINAL ANALOG MODEL

keep track of variations in the speech waveform.[1,2,4]
These parameters vary slowly.

But actually, in digital model of speech signals, the mode of excitation and resonance properties of linear system must change with time. To have a discrete time model, it is instructive to represent the various system involved in sound production with their transfer functions.

Rabiner and Schafer[1] showed that vocal transfer function can be calculated from average area functions or equivalently, reflection coefficients. An average male vocal tract was divided into N equal sections. The relation is

$$V(z) = \frac{0.5(1+r_G) \prod_{k=1}^N (1+r_k) z^{-N/2}}{1 - \sum_{k=1}^N \alpha_k z^{-k}} \quad \dots 1.4$$

where r_k is the reflection coefficient of k^{th} section

N is the no. of sections of vocal tract

r_G is the glottal wave reflection coefficient
of lossless tube,

The reflection coefficients and area functions are related as

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \quad \dots 1.5$$

where A_k and A_{k+1} are average area functions of adjacent sections of vocal tract respectively.

The vocal tract transfer functions of certain

vowels along with their area functions and reflection coefficients is shown in Figs 1.6 and 1.7.

1.2.4 Radiation

So far we considered the transfer function $V(z)$ of a vocal tract which required a configuration in which volume velocity changes occur at lips with the corresponding pressure changes at the glottis [1]. In electric analogy, medium of air is considered to be short circuit because of no impedance. The acoustic part of a short circuit is difficult to achieve as an electrical circuit because it is difficult to achieve a required configuration in which volume velocity changes will occur at lips without corresponding changes in pressure. However a reasonable model as depicted in Fig 1.8 can be made which shows the lip opening as an orifice in a sphere. At low frequencies opening can be considered as a radiating surface with the sound waves being diffracted by spherical baffle that represents the head. For determining the conditions at the lips all that is needed is relationship between pressure and volume velocity at the radiating surface. This is very complicated for the configuration of Fig 1.8(a). However if the lip opening (radiating surface) is small compared to size of sphere, a reasonable approximation assumes that the radiating surface is set in baffle of infinite extent as shown in Fig 1.8(b). It can be shown that the sinusoidal steady state relation between

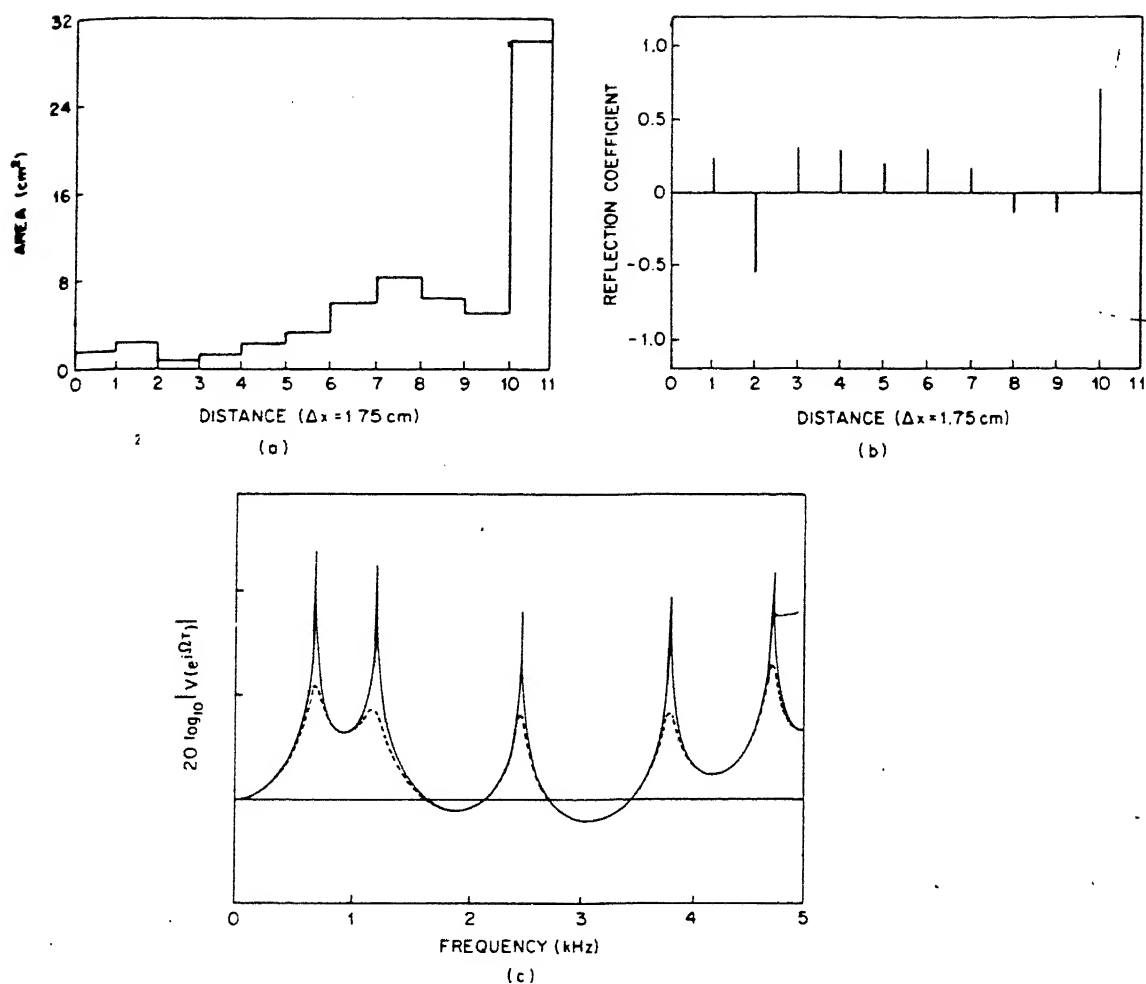
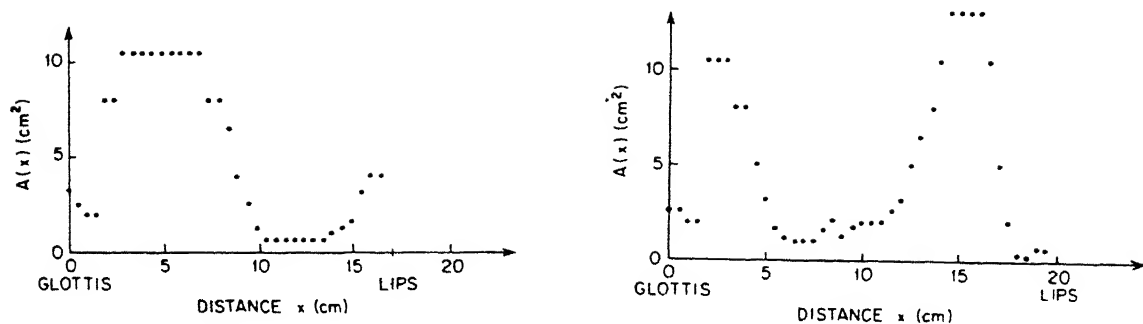
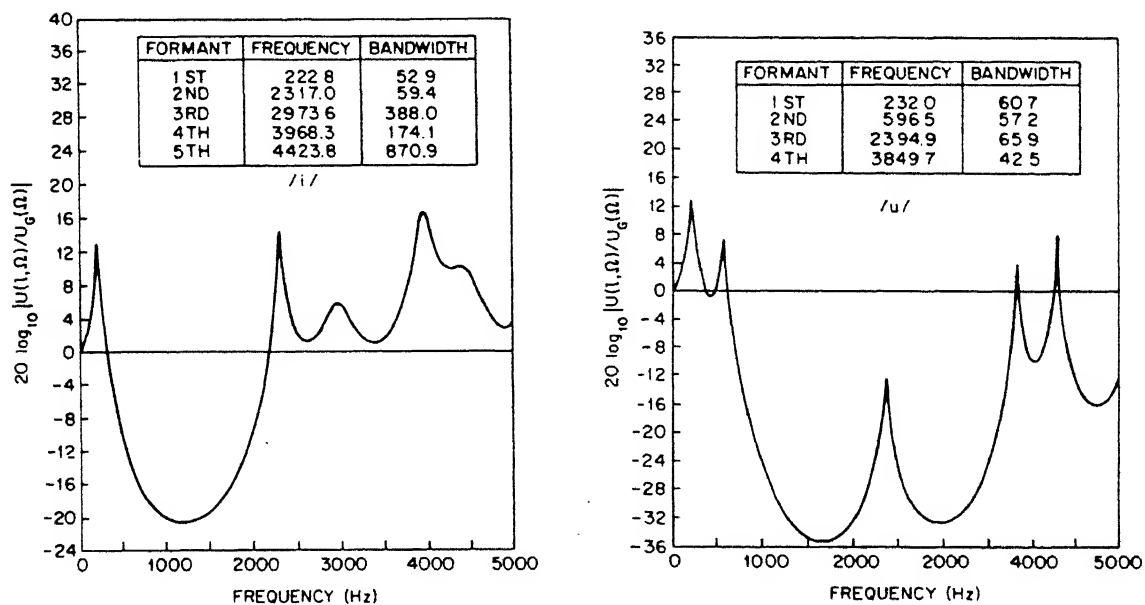


FIG 1.6(a) AREA FUNCTION FOR 10 SECTION LOSSLESS TUBE
 TERMINATED WITH REFLECTION LESS SECTION OF AREA 10 cm²
 (b) REFLECTION COEFFICIENT FOR 10 SECTION TUBE;
 (c) FREQUENCY RESPONSE OF 10 SECTION TUBE; SOLID CURVES
 CORRESPOND TO SHORT CIRCUIT TERMINATION. VOWEL: /a/.



(a)



(b)

FIG 1.7(a) AREA FUNCTIONS OF VOCAL TUBE AND
(b) FREQUENCY RESPONSE OF (a) FOR VOWEL /u/, /i/

complex amplitudes of pressure and volume velocity at the lips is through a radiation impedance or radiation load. Further, due to conversion of volume velocity into sound pressure at the microphone placed at the distant field because of differentiation of volume velocity, a genuine and comprehensive model can only be made if we consider the radiation load. Hence if we wish to obtain a model for pressure at the lips, as is the case now, the effects of radiation load must be considered. The glottal volume velocity and sound pressure for a particular vowel are depicted in Fig.1.9.[16]

1.2.5 Losses in vocal tract

So far the model we considered assumed no energy loss in vocal tube. In reality, energy will be lost as a result of

- a) viscous friction between air and the walls of tube,
- b) heat conduction through walls of tube, and
- c) vibration of tube walls.

The effects of wall vibration is caused by variation of air pressure inside the tract causing walls to experience a varying force. Thus, if walls are elastic, the cross sectional area of the tube will change depending upon pressure[1,2]. This effect results in addition of wall admittance Y in Eq 1.4 and consequent changes in reflection coefficients which were earlier real quantities.

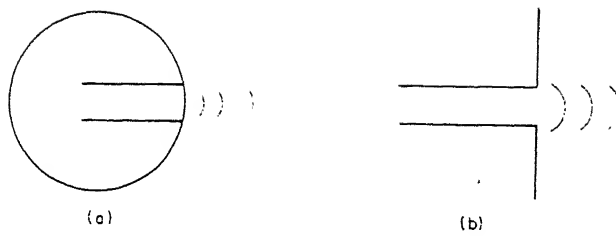


FIG 1.8 (a) RADIATION FROM A SPHERICAL BAFFLE;
(b) RADIATION FROM A INFINITE PLANE BAFFLE.

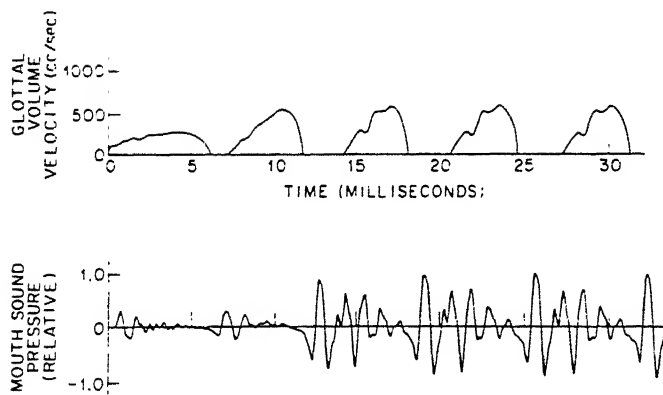


FIG 1.9 GLOTTAL VOLUME VELOCITY AND SOUND PRESSURE
AT THE MOUTH FOR VOWEL /a/.

The effects of thermal conduction and viscous friction of the wall are much less pronounced for frequencies below 3-4 KHz where as wall loss is more pronounced at these frequencies. Nevertheless, effects of wall vibration needed to be considered [6]. The model of speech mechanism however needs no alterations. Details are discussed in Chapter 5.

1.2.6 A complete model

A complete model finally can now be perceived to formulate the problem and achieve the declared objective of the thesis. Such a model is shown in Fig 1.10. It is convenient to combine the impulse train, glottal pulse, radiation and vocal tract components all together and represent them as a single transfer function $S_r(z)$ as

$$S_r(z) = P(z).G(z).V(z).R(z). \text{ or equivalantly, } \quad \dots 1.6$$

$$S_r(z) = S(z).R(z). \quad \dots 1.7$$

Here $S_r(z)$ is the transfer function of the complete model that takes into account the radiation load also. $S(z)$ is the transfer function of model which does not include the effect of radiation. $G(z).V(z)$ itself can be represented by a single transfer function $H(z)$. Thus a complete model can be represented as

$$S_r(z) = P(z).H(z).R(z). \quad \dots 1.8$$

1.3 FORMULATION OF PROBLEM

After the brief introduction to speech production

mechanism the stress of the thesis is now to approach its objective. Hence it was extremely important that the problem be formulated in a simpler way and tackled effectively.

1.3.1 Objective

Objective of the thesis is to separate the vocal tract transfer function and glottal wave transfer function by successive iteration.

1.3.2 Removal of radiation load component

It may be recalled that radiation component is a low impedance load which terminates at the vocal tract; the volume velocity of air flow at the lips (and the nose) is converted into sound pressure in distant field which is approximately the derivative of volume velocity at lips. Therefore what we wished to achieve was a model of speech for pressure at lips to retain the original identity of speech signal. To reconstruct the speech signal at lips, this radiation load component has to be removed from speech model. A sort of filter has to be applied whose transfer function will revert the influence of radiation component. Referring to Eq. 1.7, we can write the radiation load transfer function as

$$R(z) = \frac{S_r(z)}{S(z)} \quad \dots 1.9$$

In a first approximation, which is valid for lower frequencies where the wavelength is large compared to

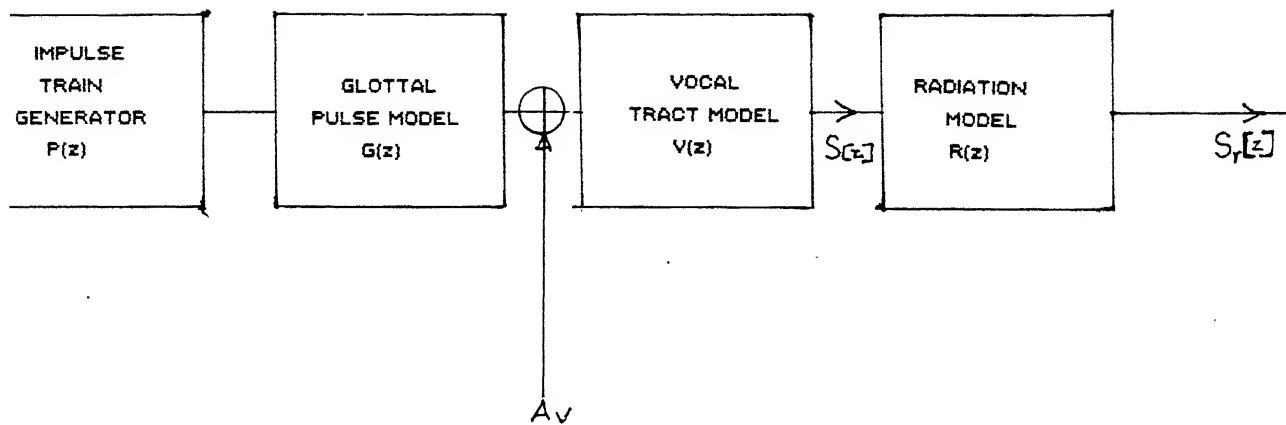


FIG 1.10 GENERAL DISCRETE TIME MODEL FOR SPEECH PRODUCTION

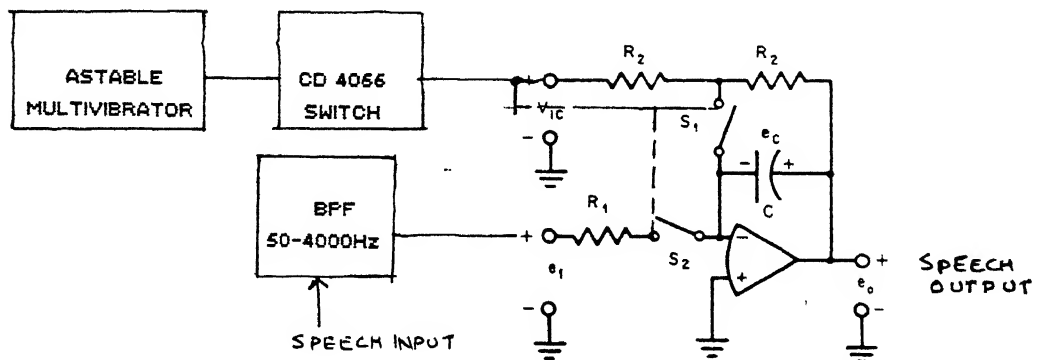


FIG 1.11 PRINCIPAL OF WORKING OF ANALOG INTEGRATOR

diameter of mouth opening, this conversion involves differentiation causing a zero at zero frequency. In the inverse filter this zero is reverted by an integrator component i.e, by a first order recursive filter with a pole at or near $z=1$. Wolfgang Hess [12] suggested the radiation load as

$$R(z) = 1 - 0.995 z^{-1} \quad \text{or} \quad \dots 1.10$$

$$S(z) = \frac{1}{1 - 0.995 z^{-1}} \cdot S_v(z) \quad \dots 1.11$$

but we used a conventional analog integrator whose performance and results differed marginally from the filter suggested by them. The speech signal output from analog integrator and inverse glottal filter for some of the vowels are shown in Figs 1.12 and 1.13 for comparison. This integrator has switching device with it to initiate and terminate the period of integration.

Fig 1.11 illustrates the functioning of this integrator [26]. In reset mode if switch S_1 is closed, initial conditions are established by placing an initial charge on the capacitor. This also allows the output voltage to rise to negative of V_{IC} . If switch S_1 is then opened and switch S_2 is closed, the circuit begins the integration of input signal e_1 beginning at the value $-V_{IC}$. An astable multivibrator coupled

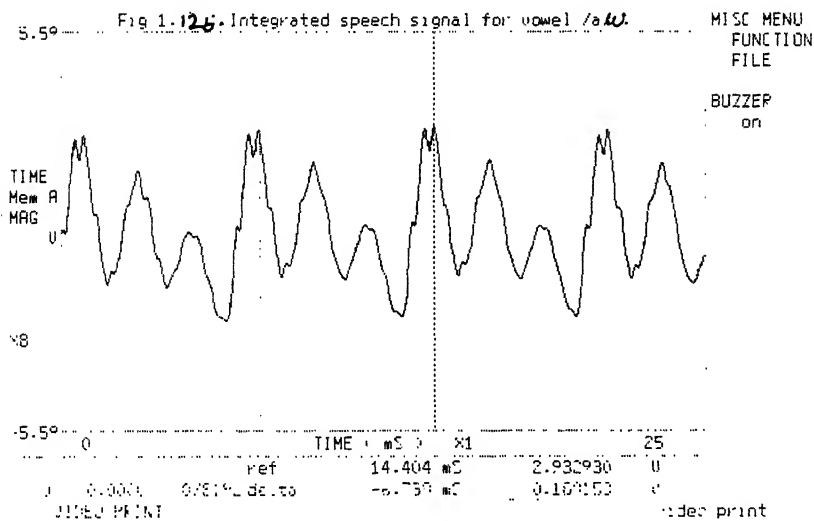
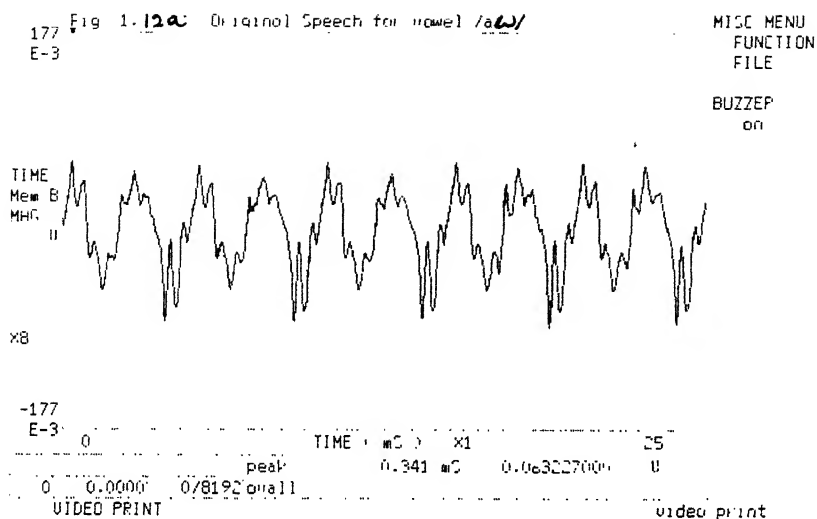


FIG 1.12(a) ORIGINAL SPEECH SIGNAL FOR VOWEL /aw/;
(b) INTEGRATED SPEECH SIGNAL OF (a) USING ANALOG INTEGRATOR.

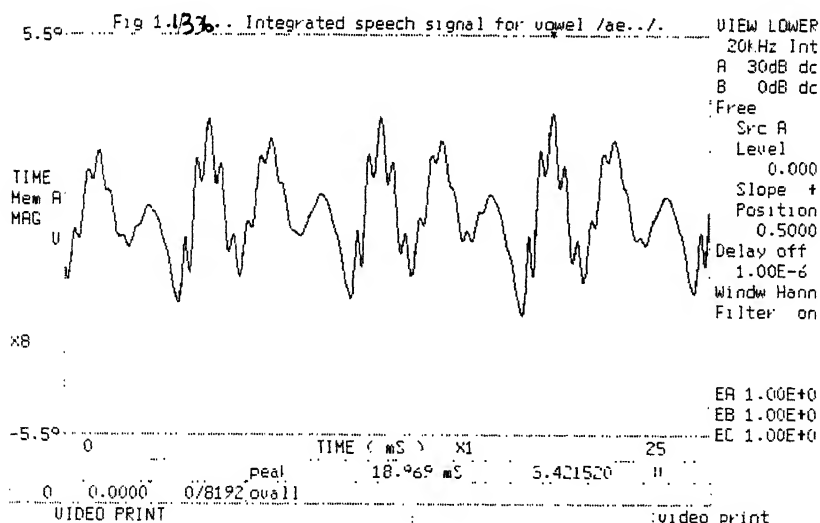
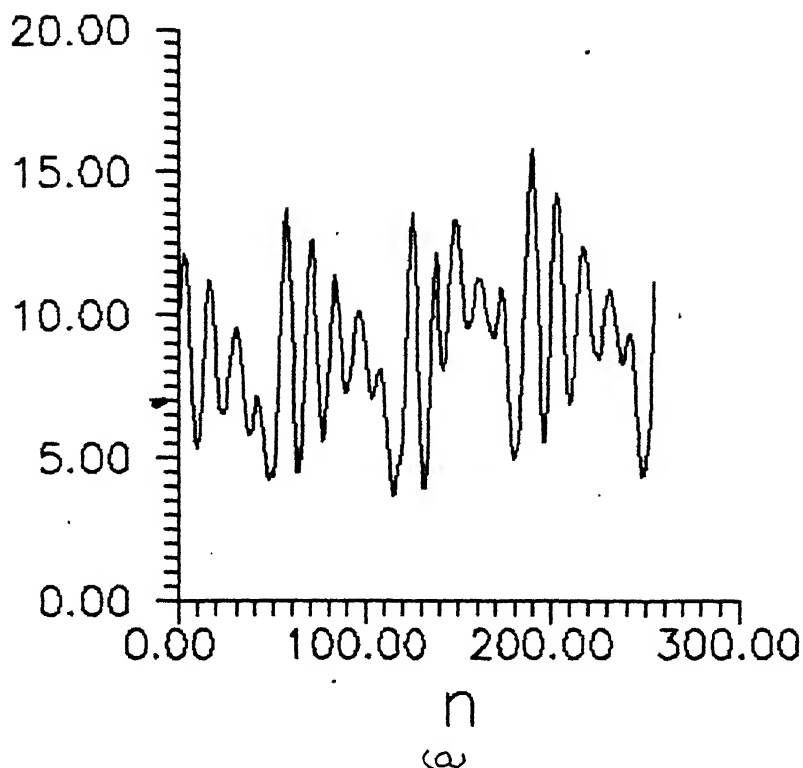


FIG 1.13(a) INTEGRATED SPEECH SIGNAL FOR VOWEL /aw/ USING DIGITAL FILTER ;
 (b) INTEGRATED SPEECH SIGNAL FOR VOWEL /ae/ USING ANALOG INTEGRATOR.

with a CD 4066 has been used as a switching device.

A band pass filter to pass 50–4000Hz frequency has been incorporated before the integrator. Speech signal was passed through this BPF to integrator. The integrated signal was then fed to FFT Analyser for its further processing.

A block diagram of hardware used and its detailed circuit is shown in Fig.1.14.

Speech signal $S(z)$ after the removal of radiation load component $R(z)$, is the product of $P(z)$ and $H(z)$. Homomorphic Deconvolution technique as discussed in Chapter 2 was applied to $S(z)$ through FFT Analyser to recover $H(z)$ from it and this $H(z)$ could then be transferred to PC through GPIB interface. The separation of $G(z)$ from $V(z)$ in $H(z)$ posed difficulties as their spectrums were superimposed on each other.

Chapter 3 contains the successive iteration method employed to find out individual area vocal tract transfer function $V_i(z)$. For average male, for certain vowel the vocal tract transfer function $V_a(z)$ is given by the Eq. 1.4. Division of $H(z)$ by this $V_a(z)$ resulted in determination of average area glottal wave transfer function $G_a(z)$. Fig 1.15 shows the wave shape of $g_a[n]$ computed by IDFT of $G_a(z)$. As is seen, waveform has high frequency components due to mismatch between individual and average area functions. The approximation of this glottal wave to the model glottal

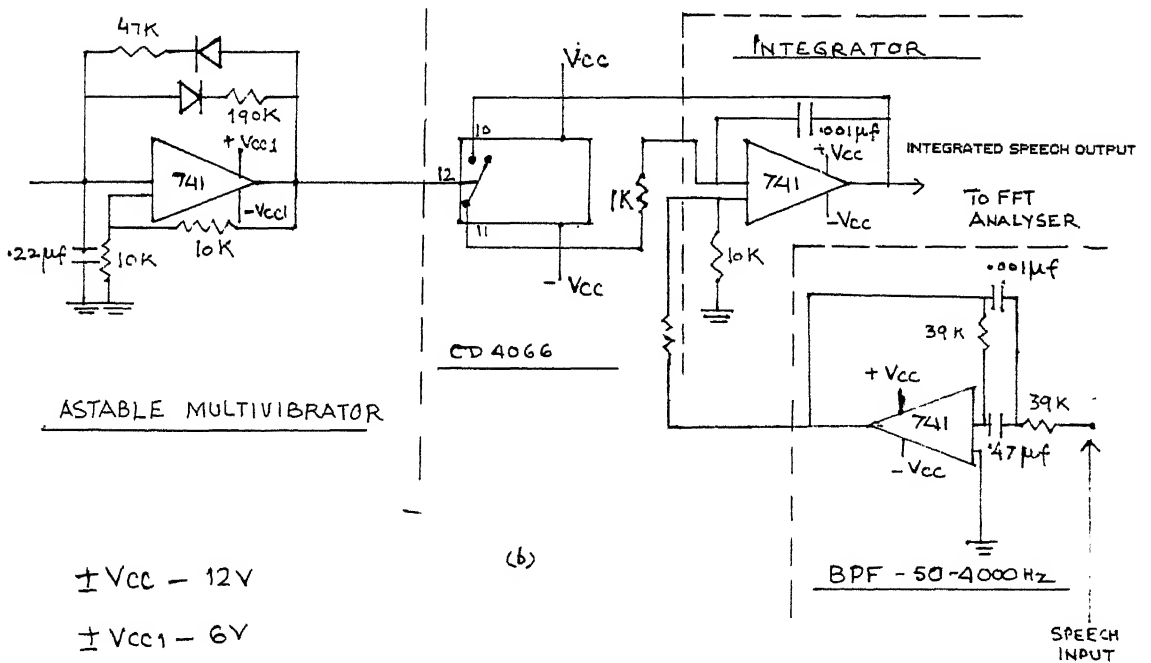
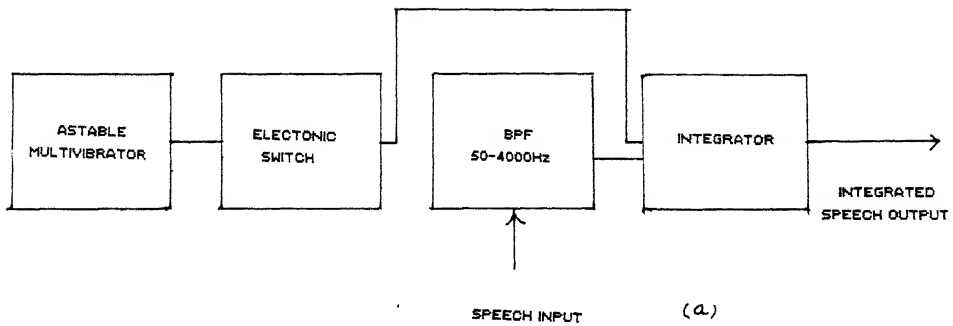


FIG 1.14(a) BLOCK DIAGRAM OF HARDWARE;
(b) DETAILED CIRCUIT DIAGRAM OF (a).

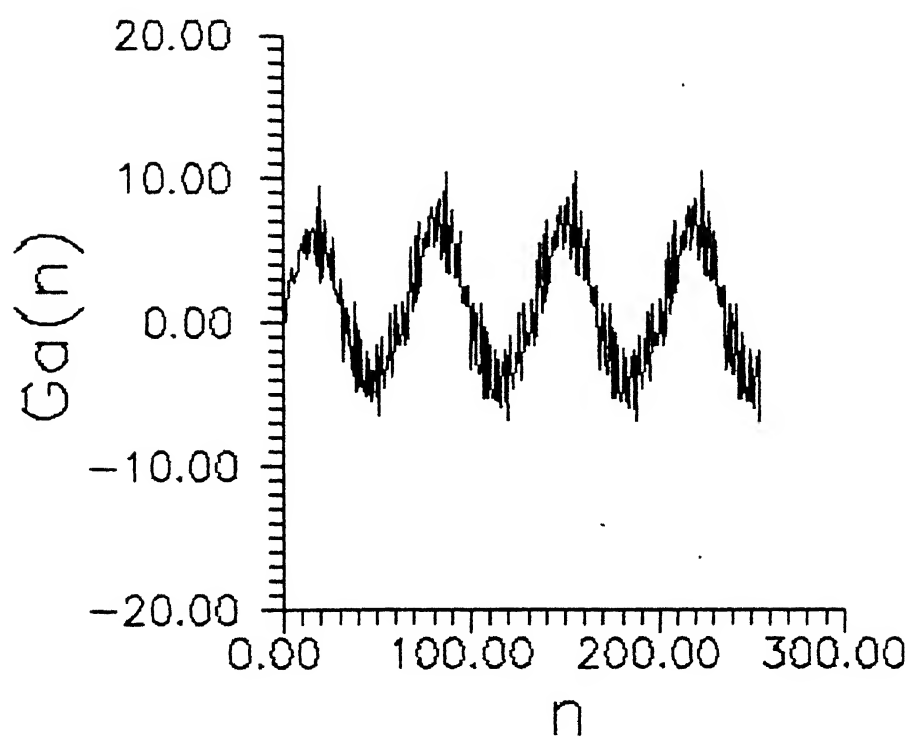


FIG 1.15 GLOTTAL WAVE SHAPE $g_a[n]$ WITH HIGH FREQUENCY COMPONENTS SUPERIMPOSED ON IT.

waveshape can be achieved either through low pass filtering of $g_2[n]$ or by finding n_2 and n_3 through Least Mean Square between individual and model glottal waveshape of Fig 1.4. The transfer function of individual glottal wave was then found out as $G_i(z)$. The resultant wave forms of $g_i[n]$ are shown in Fig 1.16. Individual area vocal tract transfer function $V_i(z)$ was found out by simple division of $H(z)$ by $G_i(z)$. This was the first step in iteration. Again this step is repeated taking this $V_i(z)$ as another $V_2(z)$ to find out another $G_2(z)$. This method is repeated several times untill most near approximation to ideal glottal wave shape is attained. The $V_i(z)$ so determined at the end of this iteration is taken as the ultimate individual $V_i(z)$.

Chapter 4 uses the relationship between linear predictor coefficients, PARCORS, and average area functions to so that individual vocal tract area functions could be calculated from this $V_i(z)$.

Chapter 5 discusses the losses that occur in the vocal tract and how they have been taken into account in $V_2(z)$ of equation 1.4.

The conclusion of the thesis is given in Chapter 6.

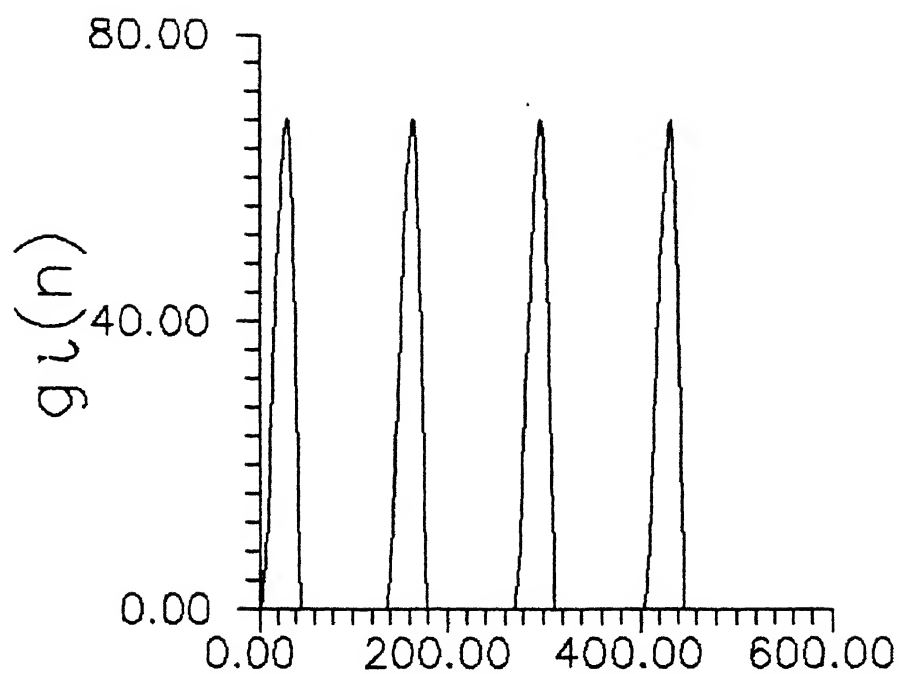


FIG 1.16 SYNTHETIC GLOTTAL WAVE SHAPE $g_v[n]$

CHAPTER 2

HOMOMORPHIC DECONVOLUTION

The fundamentals of homomorphic deconvolution is briefly given below. The details may be seen from Oppenheim and Schaffer[3].

As described in Chapter 1 there are three basic classes of sounds corresponding to different forms of excitation of the vocal tract namely, *Voiced sounds*, *Fricative/Unvoiced sounds* and the *Plosive sounds*. In each case the speech signal is produced by exciting the vocal tract system with a wideband excitation. The vocal tract changes shape rather slowly with time, and thus can be modelled as a slowly time-varying filter that imposes its frequency-response properties on the spectrum of the excitation [1,3]

Fig 1.5 depicted a discrete-time model in which the samples of speech were assumed to be the output of a time varying discrete-time system that models the resonances of the vocal tract system. Since the vocal tract changes shape rather slowly in continuous speech, it is reasonable to assume that the discrete time system in the model has fixed properties over a time interval on the order of 10 ms. (In the present case it is considered 40 ms or more because it is sound of sustained vowel with unmodulated amplitude.) Thus the discrete-time system may be characterized in each

such time interval by an impulse response or a frequency response or a set of co-efficients for an IIR system. Specifically, for voiced sounds, the transfer function of the digital filter consists of vocal component represented by

$$V(z) = \frac{A}{\prod_{k=1}^p (1 - C_k z^{-1}) (1 - C_k^* z^{-1})}$$

where p = poles of $V(z)$,

C_k and C_k^* are the complex natural frequencies of the vocal tract .

The poles $V(z)$ correspond to the formants. Thus in fig 2.1 the system function of digital filter is

$$H(z) = G(z) \cdot V(z). \quad \dots 2.2$$

This filter is excited by a train of impulses $p[n]$ in which spacing between impulses corresponds to fundamental (or pitch) period of the voice[22,24]. An amplitude control regulates the intensity of the input to the digital filter.

Homomorphic deconvolution can be applied to the estimation of parameters of the speech model if we assumed that the model is valid over short time interval[22] so that a short segment of length L samples of the sampled speech is thought of a convolution

$$s[n] = g[n] * v[n] * p[n] \quad \text{for } 0 < n < L-1 \quad \dots 2.3$$

where $v[n] * g[n]$ is the impulse response of the vocal tract system and $p[n]$ is periodic. The model of Eq. 2.3 is not

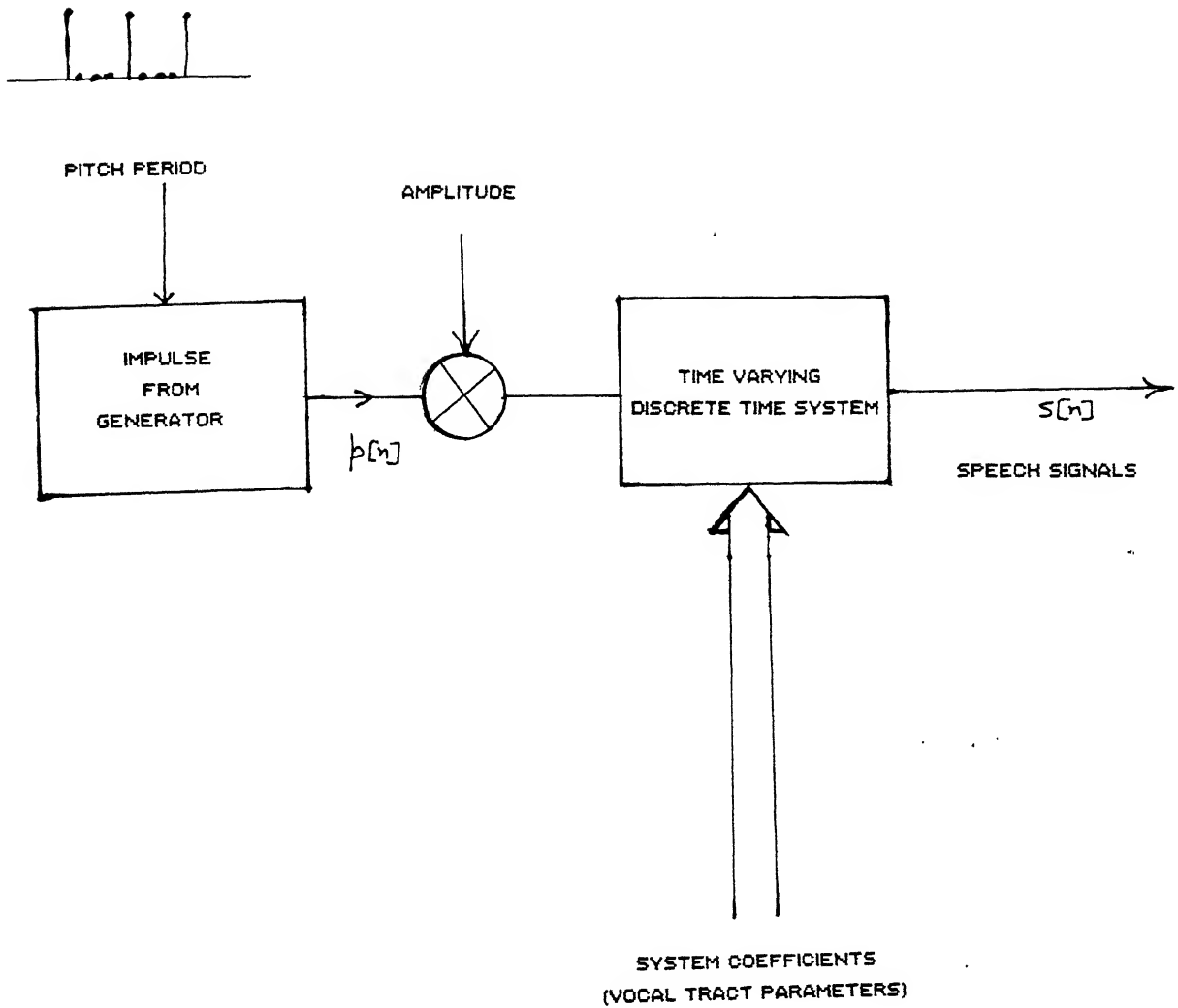


FIG 2.1 DISCRETE TIME MODEL FOR VOICED SPEECH PRODUCTION

valid at the edges of the interval because of the pulses that occur before the beginning of analysis interval and pulses that end after the end of interval. Therefore, to mitigate the effect of discontinuities of the model at the end of intervals, the speech signal $s[n]$ can be multiplied by a window $w[n]$ that tapers smoothly to zeros at both ends. Thus the input to homomorphic deconvolution system is

$$\begin{aligned} x[n] &= w[n]s[n], & w[n] &= 1 & 0 \leq n \leq N-1 & \dots 2.4 \\ & & &= 0 & n < 0 \\ & & &= 0 & n > N \end{aligned}$$

In the case of voiced speech, if $w[n]$ varies very slowly with respect to variations of $v[n]*g[n]$ [23], the analysis will be greatly simplified if we assume that

$$\begin{aligned} x[n] &= v[n]*g[n]*p_n[n] & \dots 2.5 \\ \text{where } p_n[n] &= w[n]p[n] \end{aligned}$$

Even if this assumption is not made, the detailed analysis leads us to same conclusion. [13].

Let us examine the contribution of complex cepstrum of each component of Eq 2.5. It is reasonable to assume that over short time interval of window, $p[n]$ is a train of equally spaced impulses of the form

$$p[n] = \sum_{k=0}^{M-1} \delta[n - kN_0] \quad \dots 2.6$$

where pitch period is N_0 and M periods are spanned by window.

From eq 2.6

$$p_M[n] = \sum_{k=0}^{M-1} w[kN_0] \delta(n - kN_0) \quad \dots 2.7$$

To obtain $p_M[n]$, we define a sequence

$$w_{N_0}[k] = \begin{cases} w[kN_0], & k=0, 1, 2, \dots, M-1. \\ 0, & \text{otherwise,} \end{cases} \quad \dots 2.8$$

whose Fourier Transform is (i.e, of $p_M[n]$)

$$P_M(e^{j\omega}) = \sum_{k=0}^{M-1} w[kN_0] e^{-j\omega kN_0} = w_{N_0}(e^{j\omega N_0}) \quad \dots 2.9$$

Thus, $P_M(e^{j\omega})$ and $\hat{P}_M(e^{j\omega})$ are both periodic with respect to period $\frac{2\pi}{N_0}$ and complex cepstrum of $p_M[n]$ is

$$\hat{p}_M[n] = \begin{cases} \hat{w}_{N_0}[n/N_0], & n = 0, \pm N_0, \pm 2N_0, \dots \\ 0, & \text{otherwise.} \end{cases} \quad \dots 2.10$$

where $w = w(n \cdot N_0)$.

The periodicity of complex logarithm resulting from the periodicity of the voiced speech signal is manifest in the complex speech cepstrum as impulses spaced at integer multiple of N_0 samples (pitch period) . If the sequence

$w_{N_0}[n]$ is minimum phase, then $\hat{p}_w[n]$ will be zero for $n < 0$. Otherwise, $\hat{p}_w[n]$ will have pulses spaced at intervals of N_0 samples for both positive and negative values of n . In either case, the contribution of $\hat{p}_w[n]$ to $x[n]$ will be found in the interval $|n| \geq N_0$.

The complex cepstrum of $v[n]$ can be obtained from complex logarithm of $V(z)$:

$$\hat{V}(z) = \log[A] - \sum_{k=1}^P (\log[1 - c_k z^{-1}] + \log[1 - c_k^* z^{-1}]) \quad \dots 2.11$$

From the expression it is easily seen that

$$\hat{V}[n] = \begin{cases} 0 & n < 0, \\ \log |A| & n = 0, \\ \frac{1}{n} \sum_{k=1}^P [(c_k)^n + (c_k^*)^n] & n > 0 \end{cases} \quad \dots 2.12$$

or if

$$c_k = |c_k| e^{j\phi_k}$$

$$\hat{V}[n] = \sum_{k=1}^P \frac{|c_k|^n}{n} 2 \cos \phi_k n \quad n > 0 \quad \dots 2.13$$

The glottal pulse, $g[n]$, is of finite duration and is generally assumed to be non-minimum phase sequence as in

$$g[n] = g_{\min}[n] * g_{\max}[n] \quad \dots 2.14$$

The contribution of complex cepstra $\hat{x}[n]$ due to $g[n]$ is

$$\hat{g}[n] = \begin{cases} \hat{g}_{\min}[n] & 0 \leq n \\ \hat{g}_{\max}[n] & n < 0 \end{cases} \quad \dots 2.15$$

where from our previous discussion we expect that primary contribution of $\hat{g}[n]$ to $\hat{x}[n]$ would be in the region around $n=0$.

In general, the components of complex cepstrum $\hat{v}[n]$ and $\hat{g}[n]$ decay rather rapidly, so that for a reasonable large values of N_0 , the vocal tract and glottal pulse contributions do not overlap $\hat{p}_w[n]$, that is peaks of $\hat{p}_w[n]$ stand out firmly from $\hat{v}[n]$ and $\hat{g}[n]$. In other words in the complex logarithm, the vocal tract components are slowly varying and the excitation components are rapidly varying. This is illustrated by an example.

2.1 AN EXAMPLE OF HOMOMORPHIC DECONVOLUTION OF SPEECH

Fig 2.2(a) shows a segment of speech weighted by Hamming Window [23,3]. The complex logarithm (magnitude and unwrapped phase) of the DFT of the signal in Fig 2.2(a) is shown in Fig 2.2(b). Note rapidly varying, almost periodic components due to $p_w[n]$ and slowly varying components due to $v[n]$ and $g[n]$. These properties are manifest in the complex cepstrum of Fig 2.2(c). in the form of impulses at the multiples of approximately 8 ms (the period of input speech segment) due to $\hat{p}_w[n]$ and in the region $|nT| < 5\text{ms}$ which we attribute to $\hat{v}[n]$ and $\hat{g}[n]$.

For speech sampled at 10,000 samples/sec, the

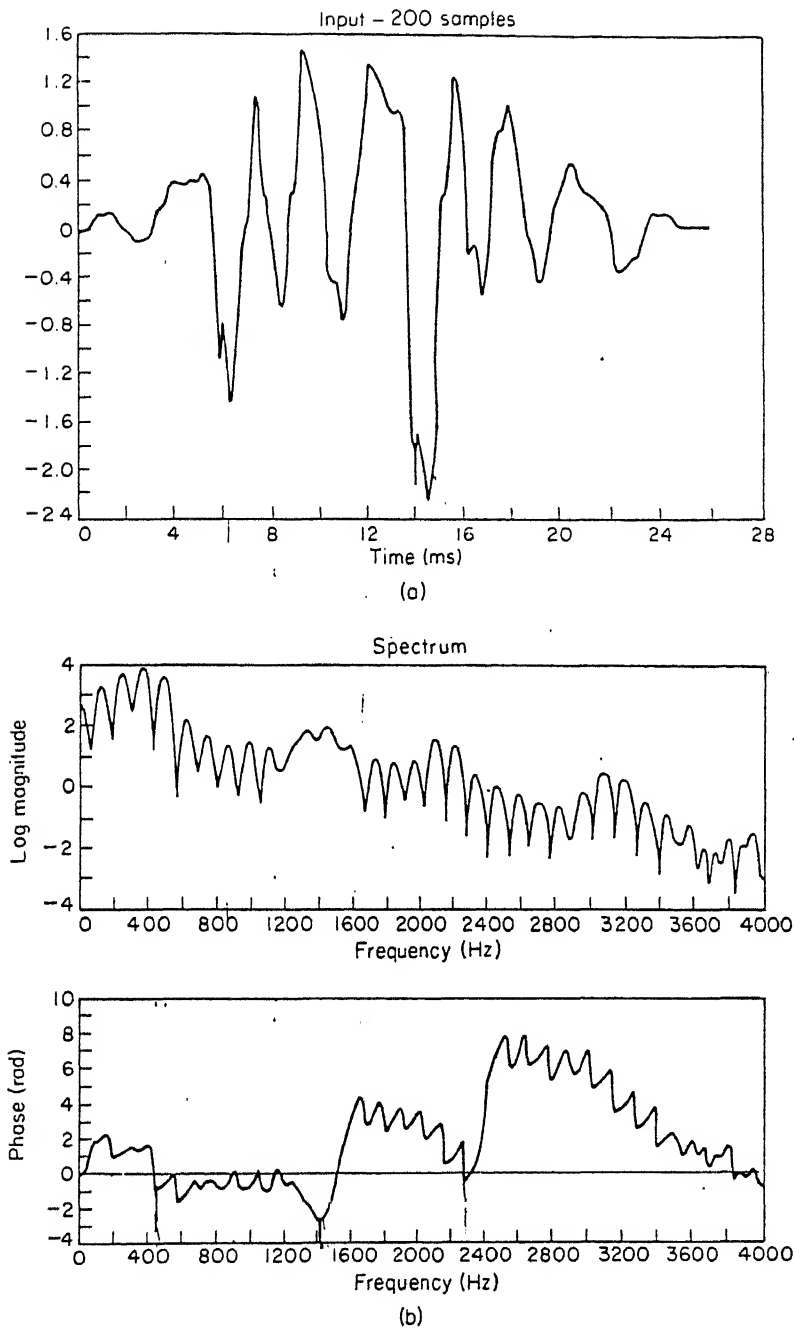
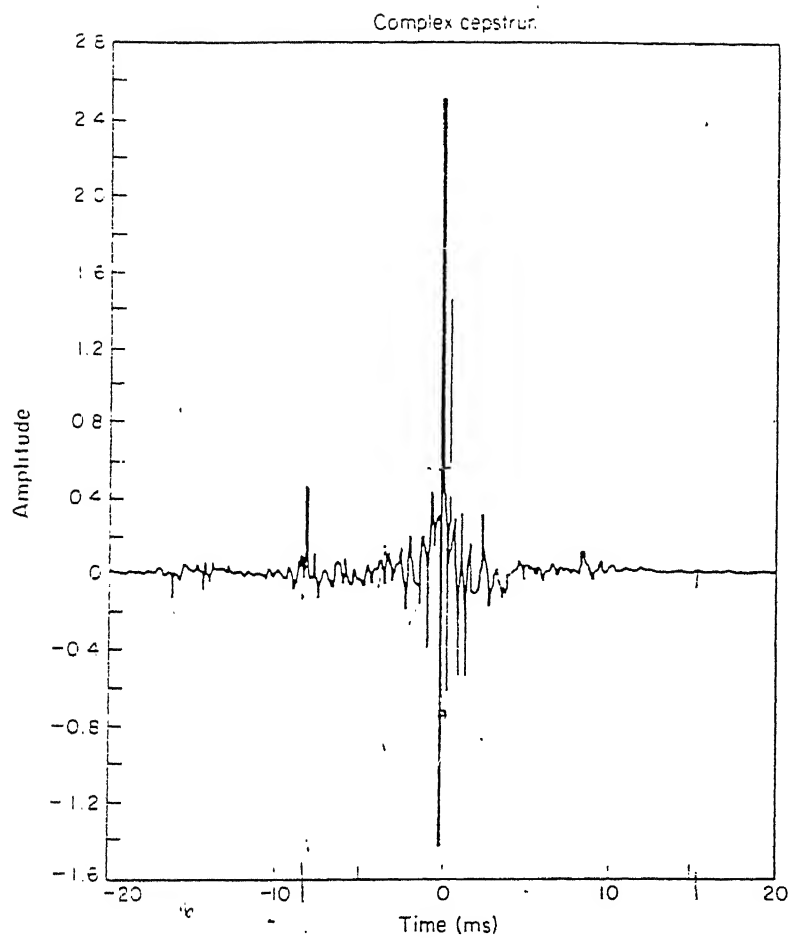


FIG 2.2(a) SEGMENT OF SPEECH WIGHTED BY HAMMING WINDOW;
 (b) COMPLEX LOGARITHM FO DFT OF SIGNAL IN (a);
 (c) COMPLEX CEPSTRUM OF PART (a).



2.2.(C)

pitch period N_0 will range from about 25 samples for high pitch voiced upto 150 samples for low pitched voice.

As previously explained, frequency invariant filter can be used to separate the components of the convolutional model of speech. Low pass filtering of complex logarithm can be used to recover the approximation to $g[n]*v[n]$ and high pass filtering for $p_w[n]$. Fig 2.3(a) show a 256 samples of vowel sound. This segment was multiplied by Hamming window and complex cepstrum was computed using DFT. The complex cepstrum is shown in Fig 2.3(b). Fig 2.3(c) is an approximation of $p_w[n]$ obtained by applying to the complex cepstrum a symmetrical high pass frequency invariant filter. Fig 2.3(d) shows approximation to $g[n]*v[n]$ obtained by using a low pass frequency invariant filter. Finally to illustrate the validity of convolution, Fig 2.3(e) shows the result of convolving the sequence of Fig 2.3(d) with an impulse train of equal amplitude impulses occurring at locations of the peaks in Fig 2.3(e). As we see by comparing Figs 2.3(a) and 2.3(e), the reconstructed waveform is very close to the original.

2.2 HOMOMORPHIC DECONVOLUTION AS APPLIED TO ACTUAL SPEECH

The previous discussion has shown that Homomorphic Deconvolution can be successfully applied for recovery of $G[z].V[z]$ from $P[z].H[z]$. A FFT Analyser in the laboratory was made use of for this purpose and the vowel uttered was 'aw'. The pitch period was 6.7 ms as shown in the markings

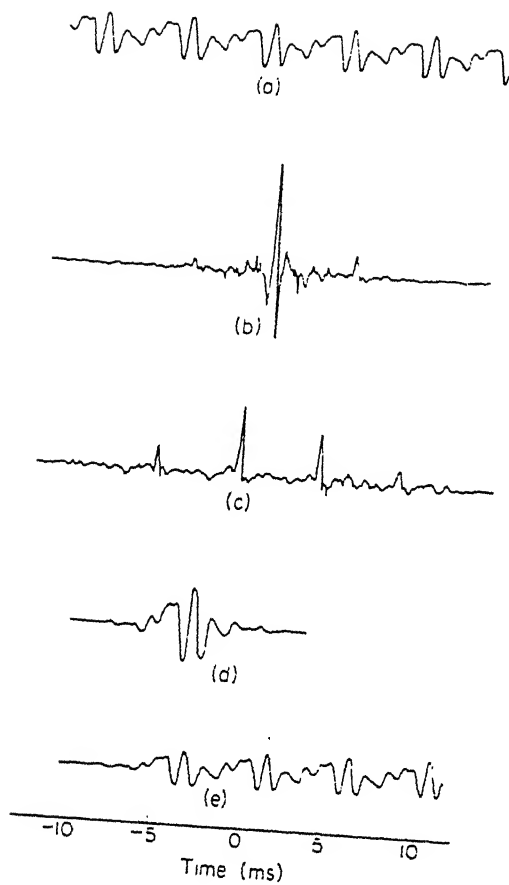


FIG 2.3(a) A SEGMENT OF VOWEL WAVEFORM; (b) COMPLEX CEPSTRUM OF PART (a); (c) RECOVERED EXCITATION FUNCTION $p_m[n]$; (d) RECOVERED $v[n]*g[n]$; (e) SYNTHESISED SPEECH USING IMPULSE RESPONSE OF PART (d) AND PITCH PERIOD AS MEASURED.

in Fig 2.4(a). The sampling frequency was 80 KHz. 536 samples were obtained from the relation

$$n = \frac{\text{pitch period}}{\text{sampling interval}} \quad \text{or} \quad 6.7 \times 10^{-3} \times 80 \times 10^3 = 536$$

Fig 2.4(b) shows the recovered complex cepstrum of the sustained vowel. This $H(z)$ was what we wanted to recover from $P(z).H(z)$. The $H(z)$ so recovered was transferred to a PC through a GPIB interface for further separation of $V_i(z)$ and subsequent determination of individual vocal tract area functions.

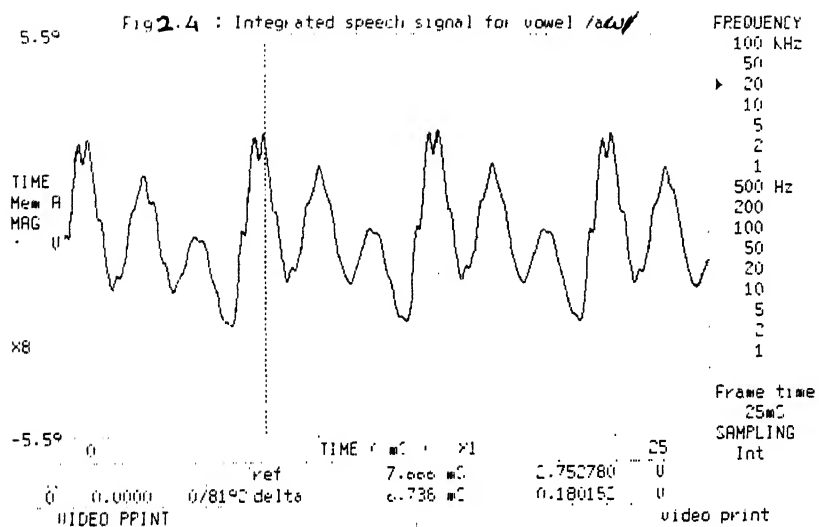
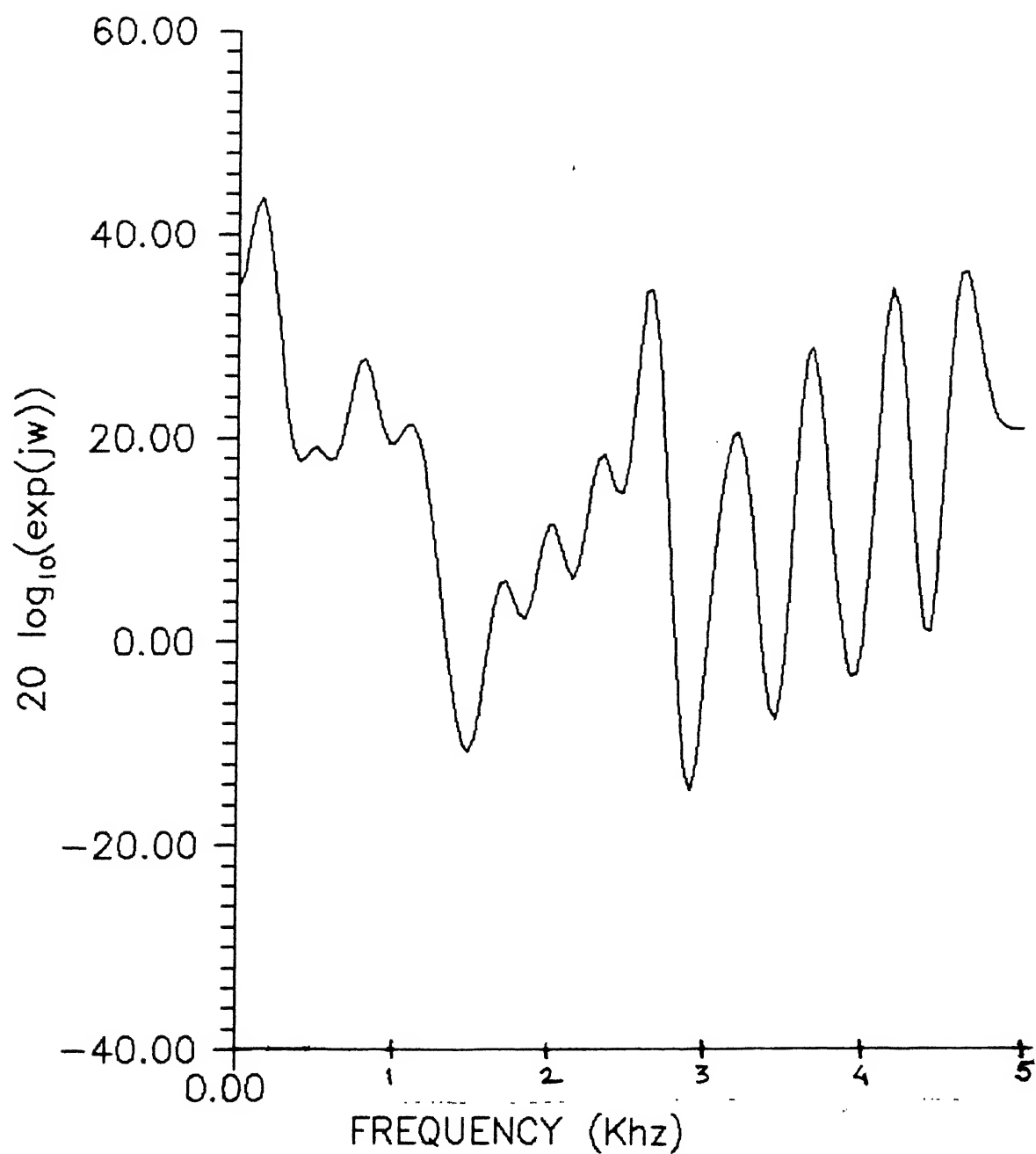


FIG 2.4(a) PITCH PERIOD OF SEGMENT OF ACTUALLY UTTERED
VOWEL /aw/

FIG 2-4(b) THE RECOVERED $H(z)$ FROM $P(z) \cdot H(z)$.

CHAPTER 3

SUCCESSIVE ITERATION

3.1 GENERAL DISCUSSION

The general expression for vocal tract transfer function as given by Eq. 1.4 was[1]

$$V_a(z) = \frac{0.5*(1+r_g) \prod_{k=1}^N (1+r_k) z^{-N/2}}{D(z)} \quad \dots 3.1$$

where

$$D(z) = 1 - \sum_{k=1}^N \alpha_k z^{-k} \quad \dots 3.2$$

is a polynomial. It assumes the lossless tube and each section to be of equal length. $D(z)$ can also be expressed as a polynomial in z^{-1} given by the matrix

$$D(z) = [1, -r_g], \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \dots \dots \begin{bmatrix} 1 & -r_N \\ -r_N z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \dots 3.3$$

In other words we say that transfer function has a delay corresponding to numbers of sections of model and has no zeros; only poles, which define resonances or formants of the model. In usual case $r_g=1$ (infinite impedance at glottis). Polynomial $D(z)$ can be found out using a recursion formula that can be derived from Eq. 3.3. The desired recursion formula becomes evident after evaluating first few matrix products. Let us define

$$P_1 = \begin{bmatrix} 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} = \begin{bmatrix} (1 + r_1 z^{-1}), & -(r_1 + z^{-1}) \end{bmatrix} \quad \dots 3.4$$

If we define

$$D_1(z) = 1 + r_1 z^{-1} \quad \dots 3.5$$

it can be easily shown that

$$P_1(z) = \begin{bmatrix} D_1(z), & -z^{-1}D_1(z^{-1}) \end{bmatrix} \quad \dots 3.6$$

Similarly, the row matrix P_2 can be defined as

$$P_2 = P_1 \begin{bmatrix} 1 & -r_2 \\ -r_2 z^{-1} & z^{-1} \end{bmatrix} \quad \dots 3.7$$

If the indicated multiplication is carried out it is easily shown that

$$P_2 = \begin{bmatrix} D_2(z), & -z^{-2}D_2(z^{-1}) \end{bmatrix} \quad \dots 3.8$$

where

$$D_2(z) = D_1(z) + r_2 z^{-2}D_1(z^{-1}) \quad \dots 3.9$$

By induction it can be seen that

$$P_k = P_{k-1} \begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & -z^{-1} \end{bmatrix} \quad \dots 3.10$$

$$= \begin{bmatrix} D_k(z), & -z^{-k}D_k(z^{-1}) \end{bmatrix} \quad \dots 3.11$$

where

$$D_k(z) = D_{k-1}(z) + r_k z^{-k}D_{k-1}(z^{-1}) \quad \dots 3.12$$

Finally, the desired polynomial is

$$D(z) = P_N \begin{bmatrix} 1 \\ 0 \end{bmatrix} = D_N(z) \quad \dots 3.13$$

Thus we can see that it is not necessary to carry out all the matrix multiplication but we can simply evaluate the

recursion

$$D_0(z) = 1 \quad \dots 3.14$$

$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}), \quad k=1, 2, \dots, N \quad \dots 3.15$$

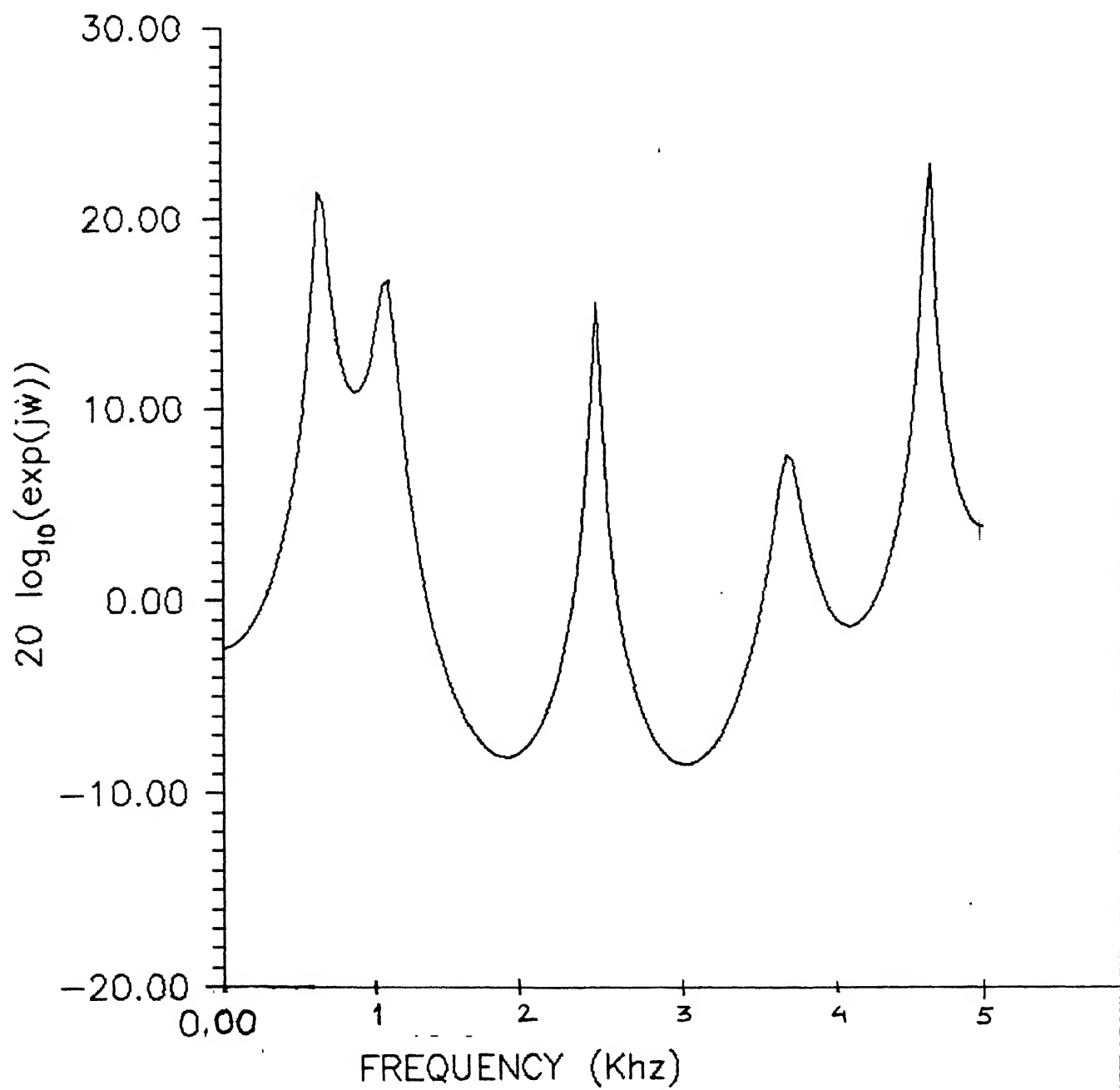
$$D(z) = D_N(z) \quad \dots 3.16$$

The effectiveness of lossless tube model can be demonstrated by computing the transfer function for the area function data given by Fig. 1.6(a).

3.1.1 The Procedure

To determine the individual vocal tract transfer function $V_i(z)$ of our speech signal, reflection coefficients r_k from the average area functions, as given in Fig 1.6 were computed using the Eq 1.5. The equation considers the radiation load of a tube of area A_{N+1} which has no reflected wave. The value A_{N+1} is chosen to give reflection coefficient at the output. This was the only source of loss in the system. Therefore r_6 was taken as 1 for infinite glottal impedance. Thus it is to be expected that A_{N+1} will control the bandwidth of the resonances of $V(z)$. Number of sections of vocal tube were taken as 10. 512 point DFT was computed from relation 3.1 using pascal language programming to obtain $V_2(z)$ of the average area vocal tract transfer function. The results of frequency response were same as depicted in Fig 1.6(c). Frequency response of our computation is shown in Fig 3.1.

The $V_2(z)$ so calculated was used for simple mathematical division of $H(z)$ of our speech signal. This

FIG 3.1 COMPUTED FREQUENCY RESPONSE $V_2[z]$

resulted in emergence of average area glottal wave transfer function $G_a(z)$. IDFT of this $G_a(z)$ provided us with digitized $g_a[n]$. This $g_a[n]$ has very high frequency components superimposed on its. This has already been shown in Fig 1.15. Our aim was to approximate this wave shape to synthetic (or modelled) glottal wave shape of Fig 1.4 and remove this high frequency components which owed their presence to mismatch between average area and individual area functions. To arrive at such an approximation, it was imperative that values of n_2 and n_3 of our individual $g_i[n]$ be calculated. These values were determined by finding the Least Mean Squared difference, between modelled $g_m[n]$ and $g_a[n]$ using the relation

$$E = \sum_{i=0}^{133} (g_m[i] - g_a[i])^2 \quad \dots 3.17$$

A range of n_2 and n_3 values were chosen for $g_m[i]$. The values of n_2 and n_3 which gave the least mean squared value E were chosen to represent synthetic $g_i[n]$ by three sets of equations 1.1, 1.2 and 1.3 given in chapter 1[27]. A point to note here is that 134 points has been chosen to represent one glottal wave cycle. This value 134 is $1/4^{\text{th}}$ of our 536 samples of our original speech data which was decimated in time. This decimation converts the original sampling frequency of 80 KHz to 20 KHz. The 512 DFT computation of $g_i[n]$ using Pascal programming produced $G_i(z)$. The transfer function $H(z)$ when divided by this $G_i(z)$,

resulted in computation of individual vocal tract transfer function $V_i(z)$.

This was the first step of our iteration method to find individual $V_i(z)$. This $V_i(z)$ then replaced the initial $V_a(z)$ and was again utilized to find another $G_a(z)$ and so on. This process of iteration was repeated several times till very near approximation to average area glottal wave transfer function was achieved.

The $V_i(z)$ which we finally got was what we desired for calculation of individual vocal tract area functions. This $V_i(z)$ is shown in Fig 3.2 as a frequency response of 10 section individual vocal tube. In chapter 4 we will deal with extraction of individual area functions from this $V_i(z)$.

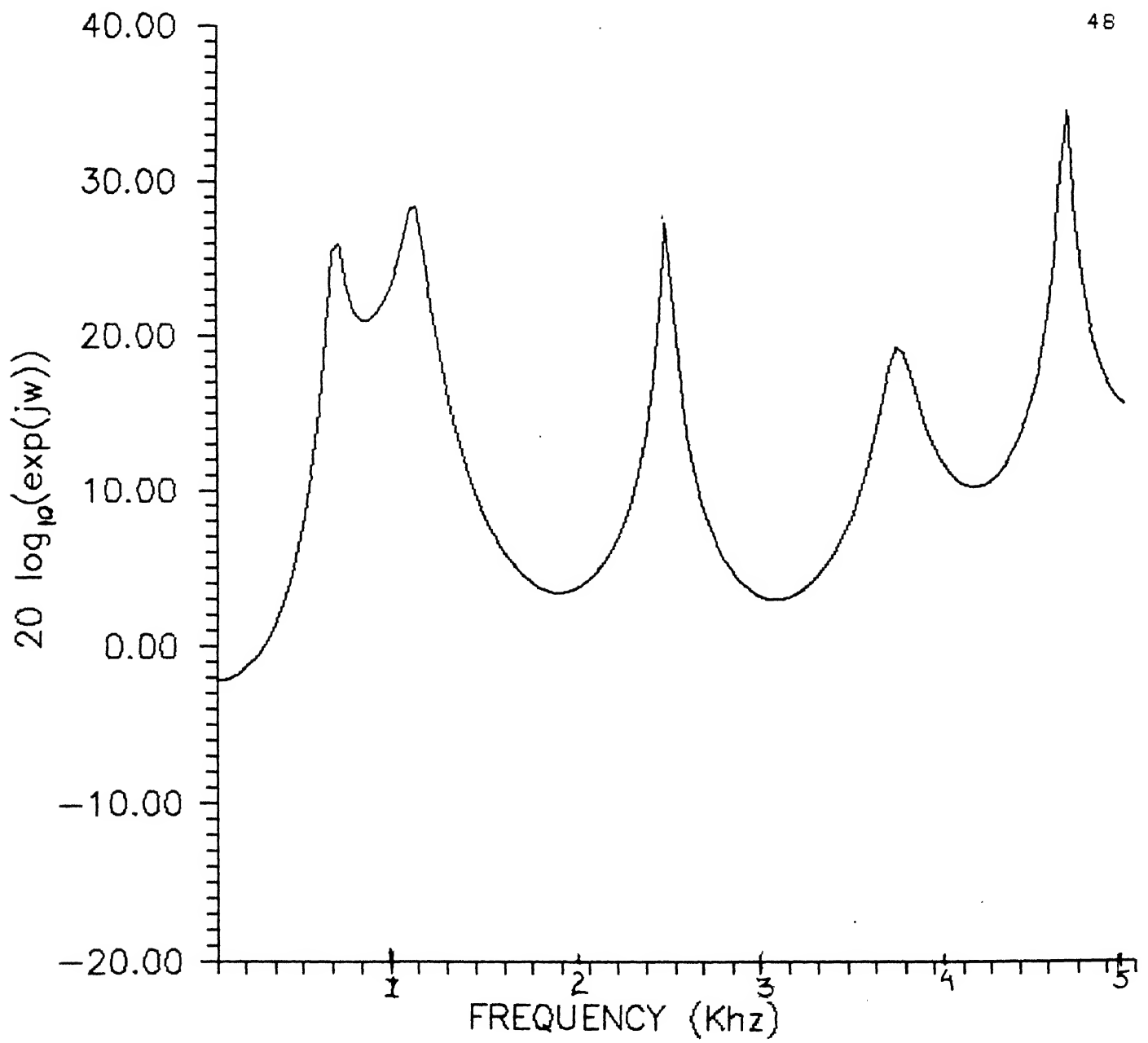


FIG 3.2 FREQUENCY RESPONSE OF INDIVIDUAL VOCAL TRACT
AS COMPUTED FOR VOWEL /aw/

CHAPTER 4

VOCAL TRACT PARAMETERS

After having obtained the transfer function for individual vocal tract, $V_i(z)$, the thesis addressed itself to extracting the individual vocal tract parameters from it. This phase required careful examination of the available techniques, their suitability, accuracy, speed of computation and reliability in our context. It was but natural that most powerful speech analysis technique was chosen. Linear predictive coding of speech signal was one such technique which met all our requirements. A brief introduction to this technique is given below.

4.1 LINEAR PREDICTIVE CODING OF SPEECH SIGNAL

4.1.1 Introduction

The method of linear predictive coding of speech signal is one of the most powerful technique for analysis of speech signals. This is the most predominant technique for estimating the basic parameters of speech like vocal tract area functions. The basic idea behind the linear predictive analysis is that a speech sample can be approximated as a linear combination of past speech samples. By minimizing the sum of squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones, a unique set of predictor

coefficients can be determined [1].

Since speech signal can be modelled as the output of linear, time varying system excited by quasi periodic pulses as in case of our sustained vowel, linear prediction method provides a robust, reliable and accurate method for estimating the speech parameters that characterise a linear time varying system. Without going into further details we will apply this technique for calculating the individual vocal tract area functions.

4.1.2 The technique as actually applied

To get the vocal tract parameters from $V_i(z)$ obtained in chapter 3, it was a must that a set of linear predictor coefficients, α_k be found out first. This $V_i(z)$ can be represented as

$$V_i(z) = \frac{1}{A_i(z)} \quad \dots 4.1$$

where $A_i(z)$ is a polynomial of the form

$$A_i(z) = 1 - \sum_{k=1}^P \alpha_k z^{-k} \quad \dots 4.2$$

obtained by linear predictor coefficients and which could also be obtained by recursion

$$A^0(z) = 1 \quad \dots 4.3$$

$$A^{(i)}(z) = A^{(i-1)}(z) - k_i z^{-i} A^{(i-1)}(z^{-1}) \quad \dots 4.4$$

$$A(z) = A^{(P)}(z) \quad \dots 4.5$$

where k_i is PARCOR coefficient. One can draw analogy between

these equations (4.3, 4.4, 4.5) and Eqs 3.14 ,3.15, 3.16. This will be more evident in section 4.2. Also Eq 1.4 and 4.1 are more intimate to each other.

4.1.3 Linear Predictor Coefficients

Although the set of predictor coefficients, α_k , $1 \leq k \leq p$ is often thought of as the basic parameters set of linear predictive analysis, it is straightforward to transform this set of coefficients to a number of other parameters set, like PARCORS to obtain vocal tract area functions [9,10]. This is how we proceed now.

The reciprocal of each point in $V_i(z)$ gave $A_i(z)$. (Eq 4.1). The IDFT of this $A_i(z)$ computed through pascal programming determined its impulse response as

$$a[n] = \delta[n] - \sum_{k=1}^p \alpha_k \delta[n-k] \quad \dots 4.6$$

where $\delta[n]$ is an unit impulse, α_k is linear predictor coefficient. The method of calculation of predictor coefficients is briefly explained as follows:

Suppose there are 10 samples in $a[n]$. Then

for $n=0$,

$$a[0] = \delta(0) - [\alpha_1 \delta(0-1) + \alpha_2 \delta(0-2) + \dots \alpha_p \delta(0-p)] = 1; \quad \dots 4.7$$

for $n=1$,

$$a[1] = \delta(1) - [\alpha_1 \delta(0) + \alpha_2 \delta(1-2) + \dots \alpha_p \delta(1-p)] = -\alpha_1 \quad \dots 4.8$$

similarly, for $n=p$, $a[p] = -\alpha_p$.

CENTRAL LIBRARY

112189
Acc No. A. 112189

..4.9

First value of predictor coefficient will be unity as illustrated above. Subsequent values of predictor coefficients came out to be negative of corresponding values of $a[n]$. (Further details may be seen in []).

The number of values taken from $a[n]$ were first 11 since we wanted to take 10 sets of predictor coefficients for our 10 section vocal tube. To make the first value a unity as per our above observations, and determine our subsequent values on the basis of this manipulation, we divided all the 11 values taken, by first value. In this way a set of 10 predictor coefficients was obtained from 2nd to 11th. The next requirement was to obtain corresponding set of 10 PARCOR coefficients from these predictor coefficients.

4.1.4 PARCOR Coefficients

Obtaining PARCOR coefficients was our next step for final results. PARCOR coefficients give us the ratio between areas of adjacent sections . A set of PARCORS could be obtained from LPC coefficients using a backward recursion of the form (see ref [1])

$$k_i = a_i^{(i)} \quad \dots 4.10$$

$$a_j^{(j-1)} = \frac{a_j^{(i)} + a_j^{(i)} a_{i-j}^{(i)}}{1 - k_i^2} \quad 1 \leq j \leq i-1 \quad \dots 4.11$$

where i goes from p , to $p-1$, down to 1 and initially we set

$$a_j^{(p)} = \alpha_j \quad 1 \leq j \leq p \quad \dots 4.12$$

In our case $p=10$. That way a set of 10 PARCORS was calculated after setting initial condition as in Eq. 4.12, i.e., 11th value was 10th PARCOR coefficient.

4.2 THE FINAL RESULT

With PARCORS obtained, the thesis was poised for final realization of objective.

Comparing Eq 3.14 and 4.5, if for a p section tube,

$r_i = -k_i$, then it is clear that

$$D(z) = A(z).$$

Using relation $r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}$ it can be shown that area of the i^{th} section of the vocal tract can be found out from

$$A_{i+1} = \frac{1 - k_i}{1 + k_i} * A_i \quad \dots 4.13$$

where A_{i+1} is the value of the area taken from Fig 1.6(a) at the 11th section of the vocal tube. [8].

From this relation (4.13), we finally achieved what we had set our sight for. All the 10 values of individual vocal tract areas and their reflection coefficients were calculated. Fig 4.1 will show the individual area functions and their corresponding reflection coefficients.

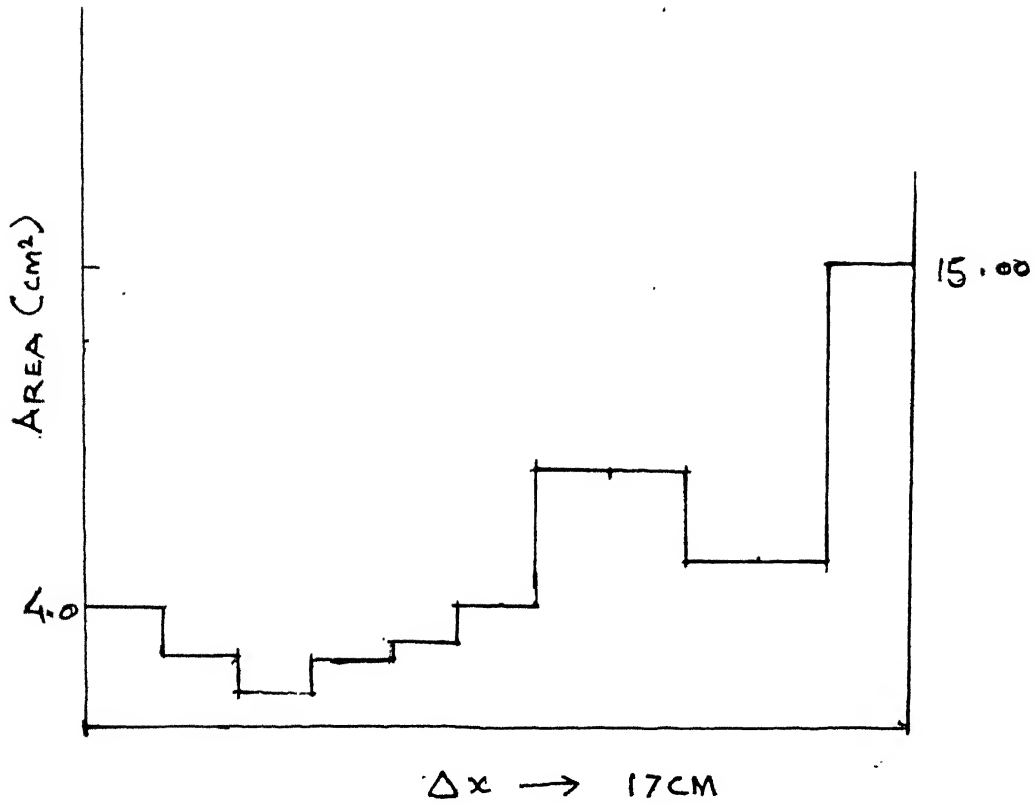


FIG 4.1 AREA FUNCTIONS OF INDIVIDUAL VOCAL TRACT
AS DETERMINED FOR VOWEL /aw/.

CHAPTER.5

LOSSES IN VOCAL TRACT

5.1 EFFECTS OF LOSSES

As had been described earlier , in chapter 1, speech coming out of lips suffers some energy losses in the vocal tract. These losses are

- (a) wall vibration loss,
- (b) viscous friction loss, and
- (c) heat conduction loss.

We shall consider each loss in some details before finally considering the total loss effected by them in the tube. To consider the effects of these losses we will be tempted to return to basic laws of motion and accoustic propogation of waves. The effects have been analysed considering the vocal tube as accoustic model.

5.1.1 Losses due to wall vibration

Because of its mass, the air in the tube exhibits an inertance which opposes acceleration. Because of its compressibility, the volume of air exhibits a compliance. The variation of air pressure inside the tract will cause walls to experience a varying force. Thus , if walls are elastic, cross sectional area will change depending upon the pressure in the tube. Assuming the walls to be locally reacting [14,15], the area $A(x,t)$ will be function of $p(x,t)$, the pressure. Since the pressure variation is small

resulting variation can be treated as small perturbation of the nominal area, i.e, we can assume that

$$A(x,t) = A_0(x,t) + \delta A(x,t) \quad \dots 5.1$$

where $A_0(x,t)$ is the nominal area and $\delta(x,t)$ is small perturbation. This is depicted in Fig. 5.1. Because of mass and elasticity of vocal tract wall [20], the relationship between area perturbation and pressure variation can be modelled by a differential equation. The details are beyond the scope of this thesis. Suffice to say the two are related by differential equations.

To consider the effect of wall vibration in frequency domain representation of a time invariant tube, excited by a complex volume velocity source, Flanagan[4] gave differential equations

$$- \frac{\delta P}{\delta x} = Z U \quad \dots 5.2$$

$$- \frac{\delta U}{\delta x} = Y P + Y_w P \quad \dots 5.3$$

where Y_w is wall admittance, Y is *acoustic admittance* per unit length of vocal tube, Z is the *acoustic impedance* per unit length of vocal tube, and U and P are frequency domain representations of volume velocity and pressure respectively. These parameters can be calculated from the

equations

$$Z(x, \Omega) = j\Omega * \frac{\rho}{A_0(x)} \quad \dots 5.4$$

$$Y(x, \Omega) = j\Omega * \frac{A_0(x)}{\rho c^2} \quad \dots 5.5$$

and

$$Y_r(x, \Omega) = \frac{1}{j\Omega m_w(x) + b_w(x) + \frac{k_w(x)}{j\Omega}} \quad \dots 5.6$$

where $m_w(x)$, $b_w(x)$ and $k_w(x)$ are mass/unit length, damping /unit length and stiffness /unit length of the vocal tract wall respectively and

ρ is the density of air in the tube,

c is the velocity of sound in the air, and

Ω is the radian frequency.

Here it is to be noted that when $A_0(x)$ is constant, the expression of Eqs 5.4 and 5.5 will reduce to

$$Z = j\Omega \frac{\rho}{A} \quad \dots 5.7$$

$$Y = j\Omega \frac{A}{\rho c^2} \quad \dots 5.8$$

which are the expressions for lossless vocal tube [7].

The effect of wall vibration is negligible for high frequencies because of very little motion of the massive walls of vocal tube [21]. But they are most

pronounced at low frequencies. Hence any vocal tube must consider this loss.

5.1.2 Viscous friction loss

Viscous friction losses are proportional to the square of particle velocity. These losses are due to viscous friction between air flow and walls of vocal tube. The effect of viscous friction losses can be accounted for in frequency domain representation (Eq 5.2 and 5.3) by including a real, frequency dependent term in the expression for the acoustic impedance, Z , [2] i.e. ,

$$Z(x, \Omega) = \frac{S(x)}{[A_0(x)]^2} \sqrt{\Omega \rho \mu / 2} + j\Omega \frac{\rho}{A_0(x)} \quad \dots 5.9$$

where $S(x)$ is the circumference of the tube,

μ is the coefficient of friction.

5.1.3 Heat conduction loss

Heat conduction losses in the smooth and hard walled tube are proportional to the square of sound pressure [2]. As in the case of viscous friction losses, the effects of heat conduction through the vocal tract can likewise be accounted for by adding a frequency dependent term to acoustic impedance $Y(x, \Omega)$; i.e.,

$$Y(x, \Omega) = \frac{S(x) (\eta - 1)}{\rho c^2} \sqrt{\frac{\lambda \Omega}{2 c_p \rho}} + j\Omega \frac{A_0(x)}{\rho c^2} \quad \dots 5.10$$

where c_p is the specific heat at constant pressure

η is the ratio of specific heat at constant pressure to that at constant volume, and λ is the coefficient of heat conduction.

Both, the heat conduction loss and viscous friction loss are only applicable to smooth, rigid vocal tube. The vocal tube is neither [18,20]. To reiterate, therefore, as we had said earlier in section 1.2.4 the effects of viscous friction and thermal conduction at the walls are much less pronounced than effects of wall vibration for low frequencies.

Having elaborated the energy losses in vocal tube, it is possible to approximate these losses or transmission characteristics as we call them in electric circuit terminology, in a π network section as shown in Fig 5.1. We shall see in the next section how all these losses can be associated with our work.

5.2 APPROXIMATION TO LOSSLESS VOACL TUBE

The total loss in the Fig 5.1 can be replaced with a single complex source Z . From this Z , the reflection coefficients for each section of the average vocal tube can be calculated from the expression

$$r_k = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i} \quad \dots 5.11$$

and from these reflection coefficients the average area transfer function $V_a(z)$ of Eq. 1.4 could be calculated and

the whole process of successive iteration as described in preceding chapters can be applied. The only difference is that the reflection coefficients will be complex quantity rather than real as obtained in Eq 1.5.

However, one may be tempted to believe that if these losses are to be accounted for, the transfer function $V_a(z)$ of average area vocal tract will be considerably different from what we had calculated earlier. This apprehension is not true. Because the yielding walls tend to raise the resonant frequencies while viscous and thermal losses tend to lower them. The net effect is slight, very slight upward shift for lower resonant frequencies, as compared to the loss less tube & rigid walled tube model. That is why the model which had been chosen by us to analyse our speech is a good representation of sound transmission in vocal tract.

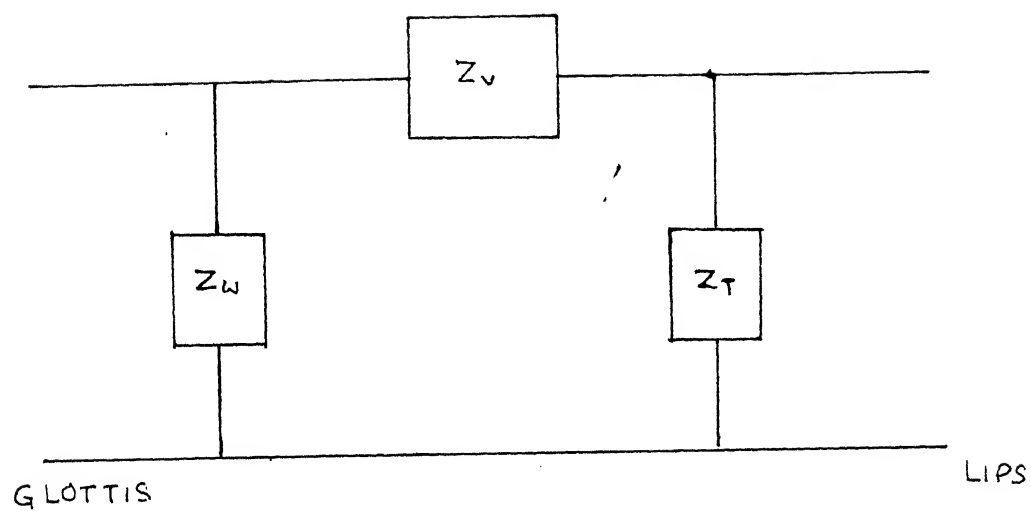


FIG 5.1 II NETWORK FOR TOTAL LOSSES IN THE VOCAL TRACT.

CHAPTER 6

CONCLUSION

The thesis has focussed its attention upon three main areas namely, the human mechanism and physics of speech production for voiced sounds, separation of glottal wave and vocal tract transfer functions from their superimposed frequency spectrum and lastly, the extraction of area functions of individual vocal tract. A discrete time model of speech was first made after the explanation of complete process of speech production including the losses. This model formed the bedrock of our thesis work.

The components of speech spectrum, $G(z)$ and $V(z)$ were separated from each other through successive iteration. The results were shown in the form of figures which have been attached with each phase of work done.

Finally, the ultimate realisation of the area function of individual vocal tract was achieved with significant success. The thesis work was tried for several sustained vowel utterances on some cases who had difficulty in uttering these vowels. The determination of the area functions of their vocal tract helped in diagnosing the disorders in their vocal tract.

SCOPE FOR FUTURE WORK

The thesis work has great scope for future work in diagnosis of disorders in any individual's vocal tract and

speakers identification and his secrecy. Presently the diagnosis of faulty speech utterances is going on ^{with} great vigour and speed. In fact this ^{is} one of the latest field in biomedical engineering where engineering skill is being amalgamated with medical science opening new vistas for the people severely handicapped by natural ^{pathological} disorders.

Another very important application of the thesis in future work is on speaker's identification (individual area function of vocal tract is "personal") and his secrecy. Enormous amount of work is going on particularly in defence research organizations on this subject. This application for defence forces needs no elaboration. The line of approach adopted is as described in the thesis.

REFERENCES

- [1] Lawrence L Rabiner and Ronald W Schafer, "Digital Processing of Speech Signals", Prentice Hall, Inc., Englewood Cliffs, New Jersey.
- [2] James L. Flanagan, "Speech Synthesis and Perception", Springer-Verlog, 2nd edition, New York, 1972.
- [3] A.V. Oppenheim and R.W. Schafer, "Digital Signal Processing", Prentice-Hall, Inc, Englewood-Cliffs, New Jersey, 1975.
- [4] G.Fant, "Accoustic Theory of Speech Production", Mouton, The Hague, 1970.
- [5] Chiba and M. Kajiyama, "The Vowel its Nature and Source", Phonetic Society of Japan, 1958.
- [6] M.M. Sondhi, "Model for Wave Propagation in Lossy Vocal Tract", J. Accoust. Soc. of Am., Vol. 55, No. 5, PP 1070-1075, May 1974.
- [7] M.M. Sondhi and B. Gopinath, "Determination of Vocal Tract Shape from Impulse Response at Lips", J. Accoust. Soc. of Am., Vol 49, No. 6, (Part 2), PP 1867-73, June 1971.
- [8] H. Wakita, "Direct Estimation of Vocal Tract Shape by Inverse Filtering of Accoustic Speech Waveforms", IEEE Trans. on Audio and Electro accustics., Vol AU-21, PP 417-427, Oct 1973.
- [9] J.D. Markel and A.H. Gray, Jr, "Linear Prediction of

Speech", Springer-Verlog, New York, 1976.

[10] J. Makhoul, "Liner Prediction: A Tutorial Review", Proc. IEEE, Vol. 63, PP 561-589, 1975.

[11] B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of 'Speech Wave", J. Accoust. Am., vol 50, PP 637-655, Aug. 1971.

[12] Wolfgang Hess, "Pitch Determination of Speech Signals", Springer-Verlog, New York, 1983.

[13] Verner Verhelst and Oscar Steenhaut, "New Model for Short-Time Complex Cepstrum of Voiced Speech", IEEE Trans on Accoust., Speech and Signal Prosessing", Vol ASSP 34, no. 1, Feb. 1986 .

[14] M.R. Portnoff, "A Quasi one Dimensional Digital Simulation for the Time Varying Vocal Tract," M.S. Thesis, Deptt. of Elect. Engg. Engr., MIT, Cambridge, Mass., June 1973.

[15] P.M. Morse and K.U. Ingard, "Theoretical Accoustics", McGraw-Hill Book Co., New York, 1968.

[16] K. Ishizaka and J.L. Flanagan, "Synthesis of Voiced Sounds from a Two Mass Model of Vocal Tract", Bell Syst. Tech. J., Vol. 50 No. 6, PP 1233-1268, July-August 1972.

[17] B.S. Atal, "Determination of Vocal Tract Shape Directly from Speech Wave", J. Accoust. Soc. Am., Vol 74, P 64, Jan 70.

[18] Man Mohan Sndhi and J.R. Resnick, "Inverse Problem for Vocal Tract: Numerical Methods, Accoustical

Experiments, and Speech Synthesis", J. Acoust. Soc. Am., Mar. 83, Vol 73, No. 3, PP 985-1002.

[19] Takiya Koizumi, Shuji Taniguchi and Seijiro Hiromitsu, "Glottal Source Vocal Tract Interaction", J. Acoust. Am., Vol 78, No. 5, PP 1541-47, Nov. 1985.

[20] Paul Milenkovic, "Acoustic Tube Reconstruction from Non-Causal Excitation", IEEE Trans. on Acoust. Speech and Signal Processing", Aug. 1987, P 1089.

[21] M. Rothenberg, "Acoustic Interaction between Glottal Source and Vocal Tract", in vocal physiology Tokyo, UP, Tokyo, Japan, PP 305 328.

[22] A.V. Oppenheim and R.W. Schaffer, "Homomorphic Analysis of Speech", IEEE Trans on Electroacoust., Vol AU 16, PP 221-226, June 1968.

[23] J.M. Tribolet, T.F. Quatieri and A.V. Oppenheim, "Short Time Homomorphic Analysis", IEEE Int. Conf. Acoust., ASSP, 1977, PP 716-722.

[24] A.M. Nole, 'Cepstrum Pitch Determination', J. Acoust. Soc. Am., Vol 41, PP 293-309, Feb. 1967.

[25] A. V. Oppenheim, "Discrete Time Signal Processing", Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[26] J. G. Graeme, G. E. Tobey, L. P. Huelsman, "Operational Amplifiers", Mc-Graw Hills, 1971.

[27] A.E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels". J. Acoust. Soc. Am., Vol 49, No. 2. PP 583-590, Feb. 1971.